Received: 6 August 2019 | Revised: 30 September 2020 | Accepted: 1 October 2020

DOI: 10.1002/widm.1394

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

WILEY

A survey of biodiversity informatics: Concepts, practices, and challenges

Luiz M. R. Gadelha Jr[1,2] | Pedro C. de Siracusa[1] | Eduardo Couto Dalcin[3] | Luís Alexandre Estevão da Silva[3] | Douglas A. Augusto[4] | Eduardo Krempser[5] | Helen Michelle Affe[3] | Raquel Lopes Costa[5] | Maria Luiza Mondelli[1] | Pedro Milet Meirelles[6] | Fabiano Thompson[7] | Marcia Chame[4] | Artur Ziviani[1] | Marinez Ferreira de Siqueira[3]

[1]National Laboratory for Scientific Computing, Petrópolis, Brazil

[2]Friedrich-Schiller-University Jena, Jena, Germany

[3]Rio de Janeiro Botanical Garden, Rio de Janeiro, Brazil

[4]Oswaldo Cruz Foundation, Rio de Janeiro, Brazil

[5]National Cancer Institute, Rio de Janeiro, Brazil

[6]Federal University of Bahia, Salvador, Brazil

[7]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Correspondence
Luiz M. R. Gadelha, Friedrich-Schiller-University Jena, Jena, Germany.
Email: luiz.gadelha@uni-jena.de

Funding information
FAPERJ; CNPq, Grant/Award Numbers: 381427/2017-0, 461572/2014-1, 441929/2016-8; CAPES, Grant/Award Numbers: 461572/2014-1, 441929/2016-8 DFG: 390713860

Abstract

The unprecedented size of the human population, along with its associated economic activities, has an ever-increasing impact on global environments. Across the world, countries are concerned about the growing resource consumption and the capacity of ecosystems to provide resources. To effectively conserve biodiversity, it is essential to make indicators and knowledge openly available to decision-makers in ways that they can effectively use them. The development and deployment of tools and techniques to generate these indicators require having access to trustworthy data from biological collections, field surveys and automated sensors, molecular data, and historic academic literature. The transformation of these raw data into synthesized information that is fit for use requires going through many refinement steps. The methodologies and techniques applied to manage and analyze these data constitute an area usually called *biodiversity informatics*. Biodiversity data follow a life cycle consisting of planning, collection, certification, description, preservation, discovery, integration, and analysis. Researchers, whether producers or consumers of biodiversity data, will likely perform activities related to at least one of these steps. This article explores each stage of the life cycle of biodiversity data, discussing its methodologies, tools, and challenges.

This article is categorized under:

Algorithmic Development > Biological Data Mining

KEYWORDS

biodiversity informatics, computational modeling, scientific data management, scientific workflows

1 | INTRODUCTION

Biodiversity is strongly related to the services that ecosystems provide, such as water, food, and climate regulation. Therefore, it is critically important to properly understand and conserve it. To meet targets on biodiversity conservation, Balmford et al. (2005) observe that it is essential to make indicators and knowledge openly available to decision-makers in ways that they can effectively use them. The Group on Earth Observations Biodiversity Observation Network (GEO BON) proposed a set of 22 Essential Biodiversity Variables (EBVs) (Pereira et al., 2013; Proença et al., 2017) that should allow for monitoring and evaluating biodiversity change. The development and deployment of mechanisms to produce these indicators depend on having access to trustworthy data from field surveys and automated sensors, biological collections, molecular data, and historic academic literature. Peterson (Peterson & Soberón, 2017), however, shows that there are information gaps across thematic and geographical areas, suggesting that there should be funding and training for institutions and personnel working on biodiversity analysis to allow for evaluating EBVs globally. The transformation of raw data into synthesized data that are fit for use requires many refinement steps. One should assess their quality (Chapman, 2005) by evaluating their taxonomic, geographical, and temporal accuracy. The methodologies and techniques used to manage and analyze these data comprise an area often called *biodiversity informatics* (Bisby, 2000; J. Soberón & Peterson, 2004; R. Guralnick & Hill, 2009; A. Hardisty et al., 2013; Hobern et al., 2013; La Salle, Williams, & Moritz, 2016). *e-Biodiversity* might also be an adequate term for describing this area if viewed as the application of e-Science (Hey & Trefethen, 2005) techniques to biodiversity.

In this survey, we give an overview of this research area covering its main concepts, practices, and some of the existing challenges. Our target audience includes researchers and students coming be applied to the challenges identified in this survey from computer science. Biodiversity data follow a life cycle (Michener & Jones, 2012) consisting of the following stages: planning, collection, certification, description, preservation, discovery, integration, and analysis. After the analysis activity, new biodiversity data management cycles may be triggered as a result. As a conceptual framework for preparing this review, we explored each step of the life cycle of biodiversity data. We grouped the steps of the life cycle in two main stages: data management and analysis and synthesis. Such steps are illustrated in Figure 1. Researchers, whether producers or consumers of biodiversity data will likely perform activities related to at least one of these steps. The remainder of this article addresses each stage of the life cycle of biodiversity data, describing their methodologies, tools, recommendations, and challenges. In the concluding remarks, we list biodiversity informatics challenges found in the review of academic literature. In particular, we build upon the Global Biodiversity Information Outlook (GBIO; Hobern et al., 2013), which maps various areas of biodiversity informatics and evaluate how much progress was achieved in each area. More recently, GBIO was used as a basis for proposing an alliance with 23 goals for developing biodiversity informatics (Hobern et al., 2019). Using these documents as guidelines, we review research in scientific databases and computational modeling for contributions that could potentially be applied to the challenges identified in this survey. In particular, in this concerns areas listed by GBIO as having minimal or limited progress.

The article is organized as follows. In Section 2, we provide some preliminary biodiversity concepts and background. In Section 3, we describe the main steps of managing biodiversity data, including the tasks involved in planning and collecting biodiversity data, data quality and fitness-for-use issues in biodiversity, the main metadata standards, and tools for biodiversity (Section 3.3), the standards, tools, and systems that support publishing and preserving biodiversity data, and techniques used to integrate biodiversity data from different sources and for discovering it. In Section 4, we discuss existing tools for analysis and synthesis of biodiversity data, including ecological niche modeling (ENM), data mining, wildlife health monitoring, network science applied to Biodiversity, Biodiversity genomics, and Biodiversity workflows management and reproducibility. Finally, in Section 5, we conclude the survey by presenting some current challenges of managing and analyzing biodiversity data.

2 | BIODIVERSITY BACKGROUND

Humanity is increasingly influencing global environments (Newbold et al., 2015). In many countries, it has raised governmental concern about the imbalance between resource consumption by human activities and the capacity of ecosystems to provide resources. This imbalance has resulted, for instance, in loss of forest cover in many places, extinction of species, and decreased availability of freshwater. Humans rely on *ecosystem services* in various activities. These services, such as food and water, are a result of processes that occur within these ecosystems. Various studies show that there is a strong relationship between human activities, global changes, biodiversity, ecosystem processes, and ecosystem

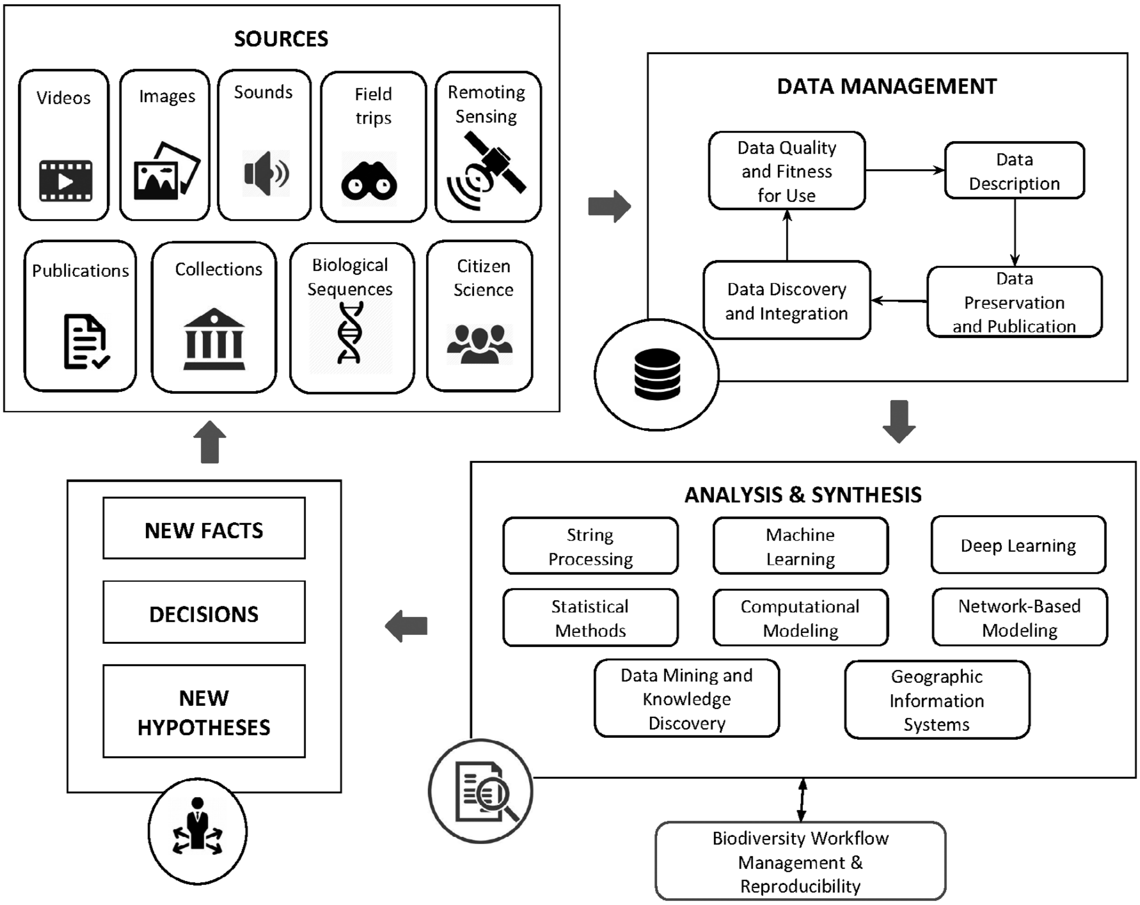


FIGURE 1 Biodiversity informatics life cycle

services (Cardinale et al., 2012; Chapin et al., 2000; Hooper et al., 2012). Chapin et al. (2000) observe that biodiversity variables, such as the number of species present, the number of individuals of each species, and which species are present, along with the interactions (e.g., trophic, competitive, and symbiotic) that are taking place between these species, determine the *species traits* that affect ecosystem processes. These traits can be defined as characteristics or attributes of species that are expressed by genes or affected by the environment. Chapin et al. (2000) also observe that global changes, often triggered by humans, such as invasive species, increased atmospheric carbon dioxide, and land-use change can significantly alter these biodiversity variables and, consequently, the expression of species traits. This, in turn, affects ecosystem processes and their resulting services, which can have negative impacts on human development. This relationship between global changes and biodiversity is illustrated in Figure 2 (adapted from Chapin et al., 2000). Changes in these ecosystem services due to changes in biodiversity can sometimes be nonlinear and stochastic, which can pose a significant risk to humans. Similar conclusions have been reached in other studies on the relationship between biodiversity, ecosystem functioning, and ecosystem services (Cardinale et al., 2012; Hooper et al., 2012). Cardinale et al. (2012) observe that after a species becomes extinct, the resulting changes to ecological processes strongly depend on which traits were eliminated. Hooper et al. (2012) observe that biodiversity loss is as significant to ecosystem change as the direct effects of global changes, such as elevated carbon dioxide in the atmosphere and ozone depletion. This, in turn, affects critical ecosystem services for the local population, such as food production, air quality, and freshwater.

A major effort to address the problem was started in 1992 during the Earth Summit in Rio de Janeiro with the signature of the Convention on Biological Diversity (CBD, 1992), a legally binding international treaty. Its main objectives are the conservation of biodiversity, including ecosystems, species, and genetic resources, and their sustainable and fair use. Countries are required to elaborate and execute a strategy for biodiversity conservation, known as a National Biodiversity Strategy and Action Plan (NBSAP), and to put in place mechanisms to monitor and assess the implementation of this strategy. They should periodically report their progress on implementing their NBSAPs. The *Strategic Plan for*

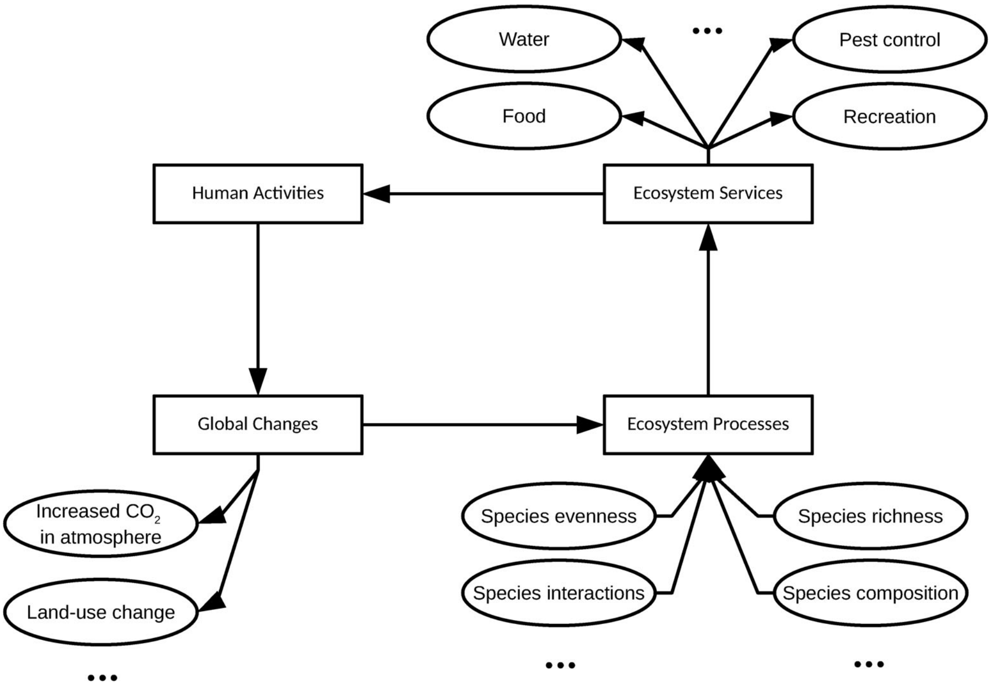


FIGURE 2 Relationship between global changes and biodiversity

Biodiversity 2011–2020 defines actions to be taken by countries to achieve a set of 20 targets by 2020, known as the *Aichi Biodiversity Targets*. It should be observed that the United Nations General Assembly declared 2011–2020 the *United Nations Decade on Biodiversity*. In 2012, the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES) was created to allow for closer cooperation between scientists and policymakers on assessing the status of biodiversity and ecosystem services and their relationship.

Throughout this work, we use definitions from the International Code of Nomenclature for algae, fungi, and plants (ICN; McNeill, 2012). This document outlines a set of rules and guidelines for scientifically naming and grouping plants, fungi, and algae, consisting of a universally adopted reference by the botanical scientific community. Nomenclature best-practices for other groups of organisms are governed by other (though similar) documents. Organism samples collected by biologists are an evidence of the existence of a particular organism at some place and time and should be properly deposited in a biological collection for being preserved as a reference. A *specimen* is defined as one such evidence and refers to a particular observation of a single kind of organism. Organisms are classified according to their shared characteristics and grouped at distinct levels of specificity (or *taxonomic ranks*) using a hierarchical system, in which groups that are more specific are nested within broader ones. The *taxonomic resolution* of a biological sample is the rank of the most specific taxonomic determination that has been assigned to it. A *taxon* is a taxonomic group of organisms at the level of any rank. *Species* is one of the taxonomic ranks in which organisms can be classified, being regarded as a basic unit of taxonomic classification. The name of a species is composed using a binomial nomenclature system, composed of the name of the genus followed by a *specific epithet*, *for example, Caryocar brasiliense*. After properly deposited in a biological collection, each record receives a taxonomic identification that assigns the individual to a *taxon*. Physical specimens stored in biological collections (also referred to as *vouchers*) are often associated with complementary information, including the *date, time*, and the *geographic location* where the specimen was collected. The taxonomic identity of a specimen includes not only the taxon name assigned to the sample, but also its nomenclatural status and authorship, the name of the person who has provided the identification. Vouchered specimens, together with their associated data, is what scientifically testifies a particular observation of a species by a collector, at some location and time, and is thus referred to as a *species occurrence*, setting up a pillar of the information used in biodiversity analysis and synthesis.

3 | DATA MANAGEMENT

From planning and collecting biodiversity data to making it fit-for-use many steps need to be followed, including data planning and collection, data quality and fitness-for-use, data description, data preservation and publication, and data discovery and integration. These steps comprise a biodiversity data management life-cycle, Figure 1 (top-right), that we present in this section.

3.1 | Data planning and collection

The various steps of the biodiversity life cycle usually comprise a Data Management Plan (DMP; Michener & Jones, 2012) of biodiversity research activities. Some research funding agencies in countries like the United States require the submission of a DMP in submissions to calls for research proposals. Many funding agencies require that proposals should include a *DMP*. A DMP is usually composed of: which data will be collected; which formats or standards will be used for these data; which metadata will be provided and in which standard or format; what are the policies for data usage and sharing; how data will be stored and how it will be preserved in the long-term; and how data management will be funded. The DMP Tool (Strasser, Abrams, & Cruse, 2014), for instance, is an online tool that supports designing and implementing a DMP. The Data Stewardship Wizard (Pergl, Hooft, Such´anek, Knaisl, & Slifka, 2019) is a web application for supporting the creation of DMPs by presenting hierarchical questionnaires to data stewards, researchers, and data scientists. These questionnaires leverage knowledge collected from the research data management community on best practices for elaborating DMPs.

Biodiversity is concerned with the variety of living organisms, which can be measured in many ways and scales, from a record of an organism observed in a geographical location at a particular date (a *species occurrence*; Yesson et al., 2007) to the relative abundance of species in a water sample collected at a long-term ecological research site (Michener, Porter, Servilla, & Vanderbilt, 2011). Omics also present many opportunities for exploring biodiversity as, for instance, molecular data from environmental samples (Robbins et al., 2012; Wooley, Godzik, & Friedberg, 2010) can be analyzed in metagenomics studies to identify functional traits and the taxonomic classification of organisms present in them. Biodiversity data can be collected in various ways: biosensor networks, field expeditions, observations made by citizen scientists, among others. In the collection process, it is important to use unique identifiers for project, sampling event, sampling area, and protocol used (Stocks, Stout, & Shank, 2016). These identifiers will later allow the data collected to be stored in biodiversity databases consistently. Whenever possible, the terms should follow a controlled vocabulary or ontology, such as the Biodiversity Collections Ontology (BCO) (Walls et al., 2014). Next, we list common sources of data that are used to describe and analyze biodiversity:

- *Species occurrences*. Species occurrences are one of the most frequently available types of data concerning biodiversity. The main attributes of a species occurrence are given by a taxon; a location; and a date of occurrence. Species occurrence records originate from different sources. To facilitate the management and improve the accessibility of such information, most institutions currently maintain it organized in digital spreadsheets or in relational database systems, while also keeping references to the physical specimens they refer to. Some institutions are even deploying efforts toward digitizing the physical specimens. Hardisty et al. (2013) observe that, at the time, only about 10% of natural history collections are digitized and that tools are required to accelerate the process. Besides specimens from biological collections, human observations are another source of species occurrences records. These observations take place, for instance, during field expeditions or even through citizen science initiatives (eBird (Sullivan et al., 2014), iNaturalist (Heberling & Isaac, 2018)). In some cases, species are maintained in the culture of living organisms, as of various collections of fungi and other organisms.
- *Species checklists*. Surveys are often performed within a geographic region, such as a continent (Ulloa et al., 2017), a country, or a national park, to determine which species are present in it. These surveys usually result in a list of taxon names called a *species checklist*. They might also be restricted to a particular kingdom or biome. Forzza et al. (2012), for instance, describe how the Brazilian Flora List, published in 2010, was assembled, which involved aggregating information about species vouchers from herbarium information systems and having taxonomists to review it. The *Catalog of Life*[1] aggregates over 100 species checklists and contained information of about 1.8 million species in 2020.
- *Sample-based and observational data*. Sample-based data are collected during events, which may be one-time or periodical, typically involve environmental data, and have a wide range and diversity of measurements. They may involve, for instance, abiotic measurements and population surveys in different temporal and spatial scales in transects, grids, and plots (Magnusson et al., 2013). They are typically collected by Long-Term Ecological Research (LTER) projects (Michener et al., 2011). Because of the heterogeneity of ecological data (Reichman, Jones, & Schildhauer, 2011), there is not a controlled vocabulary that is widely used. Some initiatives in this direction include

ontologies such as ENVO, OBOE, and BCO (Walls et al., 2014). The most common tools for publishing ecological data rely on metadata to describe tabular datasets that comprise them. Such metadata allows general information, such as dataset owner identification, geographic, temporal, and taxonomic coverages, to be recorded, facilitating their interpretation by users. Metadata also allows textually describing the meaning of each column of a tabular dataset. Later in this article, the Ecological Metadata Language (Fegraus, Andelman, Jones, & Schildhauer, 2005), a metadata standard for ecological datasets, will be described.

- *Molecular data*. The analysis of DNA, RNA, and proteins have various applications to the study of biodiversity. The genomic sequences obtained directly from environmental samples containing communities of microorganisms, that is, metagenomes (Robbins et al., 2012), for instance, provide important information to analyze their taxonomic and functional characteristics. Biological sequences can support taxonomists as well (Tautz, Arctander, Minelli, Thomas, & Vogler, 2003) in identifying species. Taxonomists can also use small genomic or gene regions to assess biological diversity across all domains of life. The Barcode of Life (Ratnasingham & Hebert, 2007; Stockle & Hebert, 2008) project, for example, analyzes and standardizes small regions of genes to help in identifying species. Some systems, such as VoSeq (Peña & Malm, 2012), allow for connecting vouchers present in biological collections to DNA sequences present in genomic databases. R. Guralnick and Hill (2009) observe that diversity can be more precisely measured, when compared to simply counting the number of species, by how species are phylogenetically related. As examples, they assess the conservation priority of North American birds using their phylogenetic distinctness and extinction risk and analyze the dispersal of the influenza A virus also using phylogenetic analysis.
- *Academic literature*. A vast amount of information about the biodiversity is available in the academic literature. Field expeditions syntheses are often available only in scientific papers. Data related to sampling, collection, and their analysis have often not been propagated to biodiversity databases. Some initiatives, such as the Biodiversity Heritage Library (BHL; Gwinn & Rinaldo, 2009), are dedicated to the digitization of historic biodiversity literature. If coupled with text extraction techniques, such as optical character recognition (OCR), one could potentially extract information from scientific articles, such as taxonomic names (Koning, Sarkar, & Moritz, 2005), and make them available in public databases.
- *Images and videos*. Field expeditions to conduct sampling often involve the production of images and videos that support the analysis of the studied sites. In the following sections, we describe Audubon Core (R. A. Morris et al., 2013), a controlled vocabulary for describing multimedia resources associated with sampling and species occurrence data.
- *Remote sensing*. According to Turner et al. (2003), most remote-sensing instruments do not have enough resolution to gather information about organisms but there were advances that enabled some aspects of biodiversity to be observed, such as differentiating species assemblages and tree species (Clark, Roberts, & Clark, 2005). They also argue that, when instrument resolution is insufficient for direct observation, indirect methods can be applied to estimate species distributions and richness. Pettorelli et al. (2016) observe that many EBVs could be derived from satellite remote sensing, which can provide global-scale regular monitoring. Some of these potential EBVs include, for instance, vegetation height and leaf area index. It is also observed that raw satellite data could be processed by scientific workflows, including tasks such as statistical analysis and classification algorithms, to generate EBVs. More recently, (Fernández, Ferrier, Navarro, & Pereira, 2020) describe the integration of on-site observations and remote sensing through biodiversity modeling for EBV estimation.

Herbarium specimens, for instance, are an important resource for documenting and analyzing biodiversity, especially its spatial and temporal patterns. However, such biological collections need to undergo a process called digitization, in which the information provided by them is converted to electronic format. In this process, a specimen can be digitally photographed and information from its labels extracted and exported to a database. This is a challenging task since there are hundreds of millions (Soltis, 2017) of specimens deposited in herbaria worldwide and processing each of them requires considerable effort. Haston, Cubey, Pullan, Atkins, and Harris (2012) describe a digitization workflow applied at the Royal Botanic Garden Edinburgh divided into three phases. The first phase involves specimen preparation, which is given, for instance, by specimen selection, movement, and taxonomic verification. In the second phase, essential data about the specimen contained in labels, curatorial records, and supplementary sources are extracted and digitized. Some of this information might also be extracted with OCR applied to label images. The third and final phase consists of capturing high-resolution images of the specimens which can be used for further information extraction with OCR, quality assessment and control, and online publication for interactive exploration by users. High throughput digitization is possible through the use of technologies such as conveyor belts and digitization stations (Borsch et al., 2020). It is also critical to keep globally unique identifiers for each digitized specimen (Güntsch et al., 2017), facilitating consistent access to information and data gathering by biodiversity portals and aggregators (Berendsohn et al., 2011). Some examples of digitization efforts include iDigBio (Paul, Mast, Riccardi, & Nelson, 2013), the Botanic Garden of Rio de Janeiro (Lanna et al., 2018), and German herbaria (Borsch et al., 2020). Some trends and future directions involve the use of artificial intelligence to automate parts of the digitization workflow, such as the automated identification of herbarium specimens using deep learning methods (Carranza-Rojas, Goeau, Bonnet, Mata-Montero, & Joly, 2017).

3.2 | Data quality and fitness for use

Although biodiversity scientists have undoubtedly benefited from open access to massive volumes of species occurrence data from many biological collections, there are some caveats that must be accounted for before using data for modeling. Data are not always adequate for investigating every aspect of natural systems, using inadequate data for studying specific aspects of biological diversity can lead to erroneous or misleading results (Chapman, 2005), and investigators must be aware of the inherent limitations of their data before formulating their questions. The availability of detailed information is still very scarce for most known organisms. This scenario, referred to as the *Wallacean Shortfall* (Lomolino, 2004), is even more critical in megadiverse countries, which still remain largely unexplored for many regions and taxonomic groups (J. Soberón & Peterson, 2004). The lack of sufficient data for threatened species is even more concerning, as designing effective programs for their conservation require knowledge on their geographic distribution and ecological requirements. This shortage of data, combined with the nonsystematic sampling and insufficient quality, limits the use of data from biological collections for many intended applications, many of which require an intensive amount of data to be available (Guisan et al., 2007). Failing to account for the inherent limitations of such data while posing and investigating their hypotheses, researchers may obtain erroneous or misleading results, eventually impacting the success of management policies that rely on such information (Chapman, 2005).

A definition for data quality based on its *fitness for the intended use* was first proposed in the context of geographical information systems (Chrisman, 1984), and became widely adopted by the biodiversity informatics community. According to this definition, quality is not an absolute attribute of a dataset but is rather given by its potential to provide users with valuable information, in specific contexts. Assessing quality attributes of data is a fundamental step for any applications that might use it, and requires that users previously delimit the purpose, scope, and requirements of their investigation. Data are considered being of high quality if it is suitable for supporting a given investigation. Depending on the application, users might need to improve the fitness of the data they have in hand, which is part of the data quality management process. Loss of quality in biodiversity data can occur during multiple stages of its life cycle (Chapman, 2005), including the moment of the recording event, its preparation before it is incorporated in the collection, its documentation, digitalization, and storage. J. Soberón and Peterson (2004) list common issues regarding biodiversity data quality. Specimens of biological collections, from which a considerable amount of species occurrence data is extracted, may have incorrect or outdated taxonomic identifications. Biological taxonomy is constantly changing to accommodate new knowledge about species. Georeferencing errors are also possible due to annotation error or instrument inaccuracy. In old records, due to the unavailability of mechanisms for accurate assessment of location, it is common to find only textual descriptions about where a specimen was collected.

It is recommended that biodiversity databases should follow as much as possible controlled vocabularies and standards for naming in order to maintain internal consistency (Chapman, 2005). Geographical coordinates, for instance, may not match the textual location description (e.g., county, state, or country name), leading to inconsistent records. When integrating data from different biodiversity databases, external inconsistencies are also a potential problem (Chapman, 2005). These happen, for instance, when names in the different databases come from lists maintained by different authorities. These may lead to missing existing links between data or even linking the data incorrectly.

One of the objectives of CBD is to establish a global knowledge network on taxonomy (A. Hardisty et al., 2013). Taxonomic concepts (Berendsohn, 1995) are often incorrectly modeled in biodiversity databases. Berendsohn (1997) developed a conceptual database model for the International Organization for Plant Information covering the different aspects and concepts that are present in taxonomy. Several tools can be used to reduce or eliminate species misidentification. For instance, official species catalogs are available online for taxon querying, such as the Catalog of Life, the World Register of Marine Species,[2] and the Brazilian Flora Species List (Forzza et al., 2012). These can be used to support taxonomic data quality assessment of occurrence records. Most of these catalogs are also accessible via application programming interfaces (APIs) available via the web, allowing the automation of this type of assessment with scripts or applications. It is important to observe that matching a name present in a biodiversity database to names in taxonomic lists does not guarantee correctness. The names may be correctly spelled according to a taxonomic list but the identification of the specimen can be erroneous. In these cases, one still needs taxonomists to check the identifications or tools that support automated identification (Carranza-Rojas et al., 2017).

Dalcin (2005) investigated data quality in taxonomic databases, proposing quality metrics and techniques for error prevention, detection, and correction using *phonetic algorithms*, such as Soundex (D. Holmes & McCabe, 2002), and *string similarity algorithms*, such as Levenshtein distance (Levenshtein, 1966). More recently, Rees (2014) observes that taxonomic names can contain errors due to misspelling which can lead to failure in retrieving data. He proposes Taxamatch, a method for approximate matching of taxonomic names. It uses a modified version of the Damerau–Levenshtein Distance (Wagner & Lowrance, 1975) algorithm for genus and species name matching and a phonetic algorithm for authority matching. Experiments showed that the method is able to identify close to 100% of errors in taxon scientific names. A. Hardisty et al., 2013 observe that there are studies about biodiversity that do not require naming organisms. For instance, metagenomic studies concentrate on analyzing samples to classify them according to functional traits identified through sequence alignment with genomic databases. For collections that have digital images of their specimens available, a promising approach is to use deep learning techniques (Schmidhuber, 2015) to automate species identification (Bonnet et al., 2018).

Regarding georeferencing problems, R. Guralnick and Hill (2009) mentions the importance of determining the georeferencing uncertainty of occurrence records and its impact on the scale at which studies can be performed. Tools like BioGeomancer R. P. Guralnick, Wieczorek, Beaman, and Hijmans (2006) and Geolocate[3] try to infer what the geographic coordinates of an occurrence of species from a textual location description. Otegui and Guralnick (2016) propose a web API that performs simple consistency checks in occurrence records, such as coordinates with zero value, disagreeing coordinates and country identification, and inverted coordinates.

Veiga et al. (2017) propose a framework for biodiversity data quality assessment and management that allows users to define their data quality requirements and when a particular dataset is fit-for-use in a standardized manner. *Data quality assessment* is given by the evaluation of fitness for use of a dataset for some application. *Data quality management* is defined as the process of improving the fitness-for-use of a dataset. The framework is given by three main components: DQ Needs, DQ Solutions, DQ Report. DQ Needs supports the definition of the intended use for a dataset, the respective data quality dimensions, acceptable criteria for data quality measurements in these dimensions; and activities to improve data quality. DQ Solutions describe mechanisms that support meeting the requirements defined in the DQ Needs component, such as tools that implement techniques to improve data quality measurements in some dimension. The DQ Report component describes the dataset that is being assessed and managed by the framework and assertions on this dataset describing measurements or amendments applied to it as specified in the other components. The authors envision a *Fitness for Use Backbone* that would implement these components and where participants could share their data quality requirements and tools. More recently, P. J. Morris et al. (2018) have extended Kurator (Dou et al., 2012), a library of data curation scientific workflows, to report data quality in terms of the data quality framework proposed by Veiga et al. (2017).

3.3 | Data description

In the description step, metadata is produced to describe biodiversity data. This metadata is essential for users to interpret datasets they download. In this section, we describe the standards, practices, and recommendations for documenting and describing biodiversity data.

3.3.1 | Ecological metadata language

The Ecological Metadata Language (EML; Fegraus et al., 2005) is a metadata standard originally developed for the description of ecological data. It is also used currently to describe datasets about species observations. The standard has several profiles with their respective fields that can be used to define the attributes of a dataset. A scientific description profile contains fields such as the creator, geographic coverage (*geographicCoverage*), temporal coverage (*temporalCoverage*), taxonomic coverage (*taxonomicCoverage*), and sampling protocol (*sampling*) used. This profile is used to define attributes of the dataset as a whole.

The data representation profile, through the *dataTable* entity, allows for describing the attributes of a tabular dataset. One can define the data types of such attributes, such as dates and numerical values, as well as their constraints, such as minimum and maximum values. Used together, the scientific description and the data representation profiles can provide good quality documentation for a dataset, supporting their meaningful interpretation. EML metadata is expressed with the XML language, illustrated in Figure 3.

Normally, biodiversity databases provide tools for editing and producing metadata in the EML standard in a more user-friendly way through a graphical interface. The DataONE (Michener et al., 2012) repository, for example, allows users to provide metadata through a graphical tool called Morpho (Higgins, Berkley, & Jones, 2002). The same repository has also a web interface called Metacat (Berkley, Jones, Bojilova, & Higgins, 2001), which allows for loading tabular ecological data in free format documented with the EML standard. The EML standard is also used to describe datasets on species occurrences and sampling events, as will be described in the following section.

3.4 | Data preservation and publication

In the preservation stage, biodiversity datasets are published in some database, such as DataONE and GBIF, where they will be available to the scientific community. These databases adopt practices of curation and management of the data aiming its preservation and availability in the long term. There are several possible procedures for publication, in this section standards and procedures for loading a dataset to a biodiversity database will be described. The publication workflow of the main current repositories will also be described.

For better management of biological collections (Schindel & Cook, 2018), several systems for this purpose have been developed in the last decades. Among the common features in this category of software are the management of specimens, control of determination history, taxonomy, images associated with specimens, bibliographic references, curatorial management activities, user management, reports tracking the evolution of collections, printing labels in varied sizes, and data quality. Among the main implementations are BRAHMS (Filer, 2013), used in more than 80 countries, allowing for working with botanical collections, Specify,[4] which is used in more than 500 institutions worldwide for more than 30 years, managing collections of flora and fauna; the Emu[5] proprietary software for managing collections, including botanical ones; and BG-Base,[6] also with more than three decades of use, widely used in botanical gardens

```
<dataset>
  <title>Baseline assessment of mesophotic reefs ...</title>
  <individualName>
    <givenName>Fabiano</givenName>
    <surName>Thompson</surName>
  </individualName>
  <organizationName>
    Federal University of Rio de Janeiro
  </organizationName>
  <abstract>Seamounts are considered important ...</abstract>
  <coverage>
    <geographicCoverage>
      <geographicDescription>Vitoria Trindade Chain</geographicDescription>
      <boundingCoordinates>
        <westBoundingCoordinate>-38.875</westBoundingCoordinate>
        <eastBoundingCoordinate>-16.375</eastBoundingCoordinate>
        <northBoundingCoordinate>-17.125</northBoundingCoordinate>
        <southBoundingCoordinate>-21.375</southBoundingCoordinate>
      </boundingCoordinates>
    </geographicCoverage>
    <temporalCoverage>
      <rangeOfDates>
        <beginDate><calendarDate>2009-03-13</calendarDate></beginDate>
        <endDate><calendarDate>2009-03-22</calendarDate></endDate>
      </rangeOfDates>
    </temporalCoverage>
  </coverage>
  ...
</dataset>
```

FIGURE 3 Part of EML metadata in XML

and arboretums. Jabot (da Silva et al., 2017) is used since 2005 in the Rio de Janeiro Botanical Garden and started to be shared in the model of cloud computing with 50 herbaria in Brazil.

3.4.1 | Darwin core

Darwin Core (DwC; Wieczorek et al., 2012) is a standard for representing and sharing biodiversity data. The standard consists of a list of terms related to biodiversity and their definitions. DwC discussions, evolution, and maintenance are conducted by TDWG (Biodiversity Information Standards), an association for the development and promotion of standards for recording and exchanging biodiversity data. DwC emerged as a term profile in the 1998 Species Analyst system developed by the University of Kansas for the management of biological collections. In 2002, it was adopted for the exchange of information in Mammal Networked Information System (MaNIS), a distributed system composed of several institutions that maintain biological collections of mammals. In 2009, the DwC standardization process was started, which was ratified in October of the same year at the TDWG annual meeting. DwC is based on the Dublin Core[7] standard, taking advantage of its terms for resource description such as *type*, *modified*, and *license*, and complementing them with specific biodiversity terms, such as *catalogNumber* and *scientificName*.

The DwC vocabulary terms are organized as follows. The *classes* indicate the categories or entities defined in the standard. Examples of classes are: *Event*, *Location*, and *Taxon*. Each class has a set of *properties*, which are its attributes. For example, the *Location* class has attributes such as *country* and *decimalLatitude*. Finally, values can be assigned to properties, such as "Chile," −33.61 for the *country* and *decimalLatitude* properties, respectively. It is worth noting that it is recommended that, whenever possible, the values come from some controlled vocabulary, in the case of textual values, or some formatting standard, in case of numerical or temporal values. For example, species names from some recognized list of species, such as the Catalog of Life. Table 1 illustrates the representation of species occurrence data with DwC. These records come from a dataset published through the Brazilian Marine Biodiversity Database (BaMBa; Meirelles et al., 2015) in GBIF.[8]

Normally, a dataset in the DwC format is accompanied by metadata, which is defined in the EML standard (Fegraus et al., 2005). In EML, fields such as the title, authors, geographic, and temporal coverage of the dataset are found, which help users interpret datasets formatted in the DwC standard.

Like relational databases, datasets that follow the DwC format can contain multiple tables that are related through attributes that are common among them. Such an organization allows, for example, sample data to be expressed also in this standard. Tables 2 and 3 illustrate this type of data organization to represent species sampling. Table 2 contains the sampling events, four in total. The *eventId* column contains an identifier for each event. The other columns describe the event date, latitude, and longitude, respectively. Table 3 contains counts of organisms for each event. The *eventId* column, describes which event in Table 2 the counts refer to. For example, the first two rows in the table refer to the event that has identifier 1, which is associated with a sampling performed on March 18, 2009.

TABLE 1 Species occurrences represented with DwC

occurrenceID	***eventDate***	***decimalLatitude***	***decimalLongitude***	***scientificName***
UCSC:VTCFishes:1	2002-08-01	−20.805828	−37.761231	*Acanthocybium solandri* (Cuvier, 1832)
UCSC:VTCFishes:109	2002-08-01	−22.382222	−37.587500	*Balistes vetula* (Linnaeus, 1758)
UCSC:VTCFishes:178	2002-08-01	−19.848744	−38.134635	*Thunnus atlanticus* (Lesson, 1831)
UCSC:VTCFishes:210	2002-08-01	−20.525417	−29.310350	*Rhomboplites aurorubens* (Cuvier, 1829)

TABLE 2 Sampling events represented with DwC

eventId	***eventDate***	***decimalLatitude***	***decimalLongitude***
1	2002-08-01	−20.805828	−37.761231
2	2002-08-01	−22.382222	−37.587500
3	2002-08-01	−19.848744	−38.134635
4	2002-08-01	−20.525417	−29.310350

TABLE 3 Species occurrences related to the events in Table 2

eventId	*organismQuantity*	*scientificName*
1	1	*Holacanthus ciliaris* (Linnaeus, 1758)
1	3	*Lutjanus vivanus* (Cuvier, 1828)
2	2	*Lepidocybium flavobrunneum* (Smith, 1843)
…	…	…

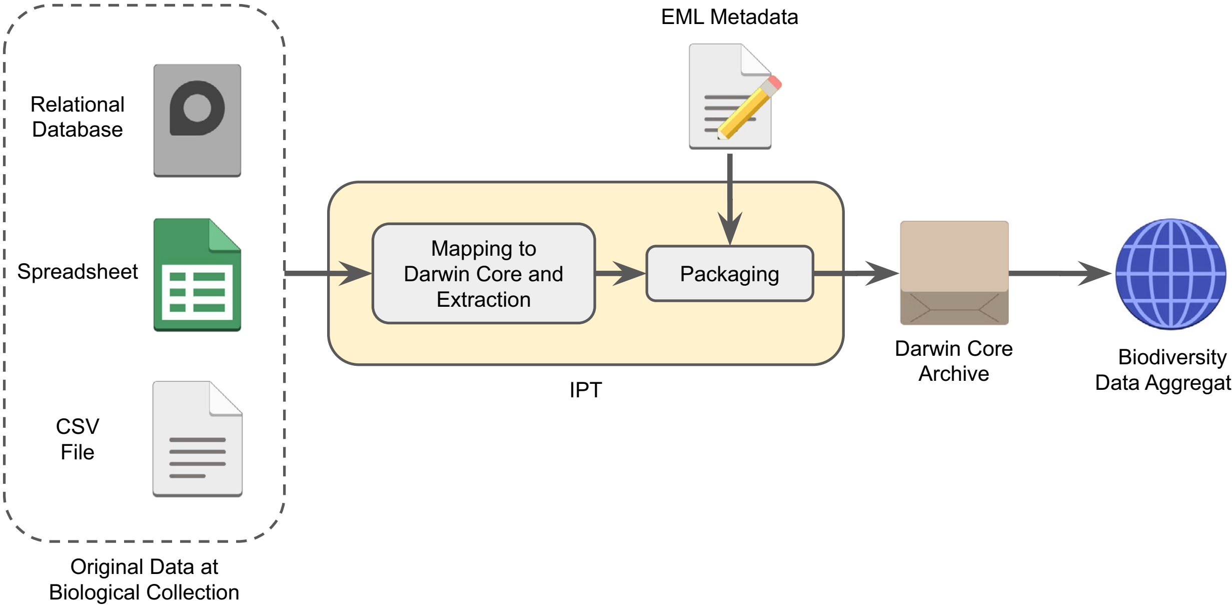

EML Metadata
Relational Database
Spreadsheet
CSV File
Original Data at Biological Collection
Mapping to Darwin Core and Extraction
Packaging
IPT
Darwin Core Archive
Biodiversity Data Aggregator

FIGURE 4 Data publication using integrated publishing toolkit (IPT)

Biodiversity datasets formatted with DwC can be published in global-scale biodiversity databases such as GBIF (Edwards, 2000). GBIF acts as a central registry and aggregator for datasets published by its national and organizational nodes using the *Integrated Publishing Toolkit* (IPT; Robertson et al., 2014). The publishing workflow includes the following steps: (i) mapping the internal representation of biodiversity data to DwC and extracting them; (ii) adding EML metadata describing biodiversity data; (iii) packaging both EML metadata and DwC-formatted data to a *Darwin Core Archive* (DwC-A); (iv) GBIF and national biodiversity aggregators, such as the Brazilian Biodiversity Information System (SiBBr) (Gadelha et al., 2014), harvest DwC-A and ingest them into their databases. This process is illustrated in Figure 4.

3.4.2 | Other data publication workflows

Many research groups have tabular biodiversity data stored in various formats and do not have the resources to format them according to DwC. Different approaches in ecology, coupled with distinct research traditions, both in their subdisciplines and in related fields, lead to the production of highly heterogeneous data. Such data can be, among others, counts of individuals, measures of environmental variables or representations of ecological processes. The terminologies used also vary according to the research line, as well as how to structure the data digitally (Jones, Schildhauer, Reichman, & Bowers, 2006). The EML metadata standard was also adopted for describing the ecological datasets. The datasets themselves, due to the heterogeneity, are published in the original format, through spreadsheets or textual files with values separated by commas. In these cases, Metacat (Berkley et al., 2001) can support data publication and preservation. It is responsible for receiving, storing, and disseminating datasets of, for example, Long-Term Ecological Research (LTER; Michener et al., 2011). The Brazilian Marine Biodiversity Database (BaMBa; Meirelles et al., 2015), for instance, was developed to store large datasets from integrated holistic studies, including physicochemical, microbiological, benthic, and fish parameters. BaMBa is linked to SiBBr and has instances of both IPT and Metacat, making it possible to publish data using the workflows previously described.

The publication of data, and its consequent preservation, is a contribution to the scientific community as a whole. The data can be reused by other scientists who can explore them from other points of view. For the publisher of the data, benefits can also be observed. A recent study (Piwowar & Vision, 2013) shows that articles that provide the data used in their analyses in public repositories tend to have a larger number of citations. Data publication and preservation can also be helpful when biological collections are lost due to disasters. On September 2, 2018, there was a catastrophic fire in the Brazilian National Museum, destroying the vast majority of approximately 20 million items in its collections comprising areas such as archeology, anthropology, zoology, and botany. Many destroyed items belonged to biological collections, including one on invertebrates. Through data publication on GBIF,[9] the museum was able to preserve information about many specimens, 269,660 records were available on September 20, 2018, many of these containing images.

3.5 | Data discovery and integration

The search for data to perform biodiversity analysis and synthesis research is still a challenging task. The most recent developments have occurred with the emergence of databases that aggregate datasets at global and national scales such as GBIF (Edwards, 2000), DataONE (Michener et al., 2012), SiBBr (Gadelha et al., 2014), and speciesLink network (Canhos et al., 2015). The use of metadata and data publishing standards allows institutions to map the internal representations of this information to a format that is clearly specified and can be consumed and processed automatically by machines. Biodiversity information aggregation databases allow datasets to be geographically, taxonomically, and temporally searched. Languages and data analysis environments, such as R and Python, already have packages and libraries that are integrated with the repositories and aggregators of biodiversity data. R*gbif*,[10] for example, is a package for R that allows searching and retrieving records directly from GBIF, with *pygbif*[11] being its analog for Python.

Often scientists need to combine data from different sources into integrative research. For example, physicochemical data can be combined with metagenomic data to try to establish correlations that explain some ecosystem processes and their implications to the effectiveness of marine protected areas (Bruce et al., 2012; Meirelles et al., 2015). The activity of combining data from different sources is called data integration and is one of the most active areas of research on scientific data management (Ailamaki, Kantere, & Dash, 2010; König et al., 2019; Miller, 2018). Existing biodiversity databases have advanced by establishing standards for metadata, such as EML (Fegraus et al., 2005), and for data such as DwC. However, these are limited to defining controlled vocabularies, consisting of standardized terms in each of the themes. A more sophisticated approach, involving not only the definition of terms, but also the relationships between them and rules of inference, which are called ontologies, is the subject of the *Semantic Web* research area. Some initiatives in this direction in the area of biodiversity and ecology include ontologies such as the *Environment Ontology* (ENVO) and the BCO (Walls et al., 2014). Ontologies allow cross-referencing of different domains (*Linked Data*) and semantic queries, providing a data integration tool considerably more powerful than the current ones.

4 | DATA ANALYSIS AND SYNTHESIS

In this section, we present some examples where biodiversity data is analyzed along with the computational methods used. These main biological examples are related to ecological niche modeling (ENM), network science, biodiversity genomics, wildlife health monitoring, and biodiversity data mining. We also explore methods for interconnecting various computational tasks, that is, biodiversity workflow management, and for keeping track of data derivation in these workflows with the purpose of enabling analysis and synthesis reproducibility, which are essential in managing biodiversity analysis and synthesis activities.

4.1 | Ecological niche modeling

ENM is used to predict the potential geographic distribution of a given species based on environmental factors (Peterson et al., 2011). A niche-based model represents an approximation of the fundamental ecological niche of a species in the environmental dimensions analyzed (Peterson et al., 2011; Phillips, Anderson, & Schapire, 2006). This model is made using a family of statistical tools to analyze the environmental information associated with the occurrence points (geographic coordinates), generating maps with an indication of geographic areas with the environmental suitability of the modeled species (Elith & Leathwick, 2009; Gomes et al., 2018). Different studies include ENM with ecological and evolutionary objectives of analyzing and is increasingly incorporated in decision-making, for example, the potential distribution of invasive species (e.g., [Peterson & Robins, 2003]), with an indication of vulnerable areas, the distribution of species in scenarios of climate change (e.g., M. Araújo, Nogués-Bravo, Reginster, Rounsevell, & Whittaker, 2008; M. B. Araújo & Peterson, 2012; Pearson, Thuiller, et al., 2006; Thomas et al., 2004; Wiens, Stralberg, Jongsomjit, Howell, & Snyder, 2009), dissemination of infectious diseases (e.g., [Costa, Peterson, & Beard, 2002]), and selecting conservation areas (e.g., [M. B. Araújo & Williams, 2000; Y. Chen, 2009; Engler, Guisan, & Rechsteiner, 2004; Pearson, 2010]). The concept of niche is defined, according to Chase and Leibold (Chase & Leibold, 2003), as environmental conditions that meet the minimum requirements of a species so that its birth rate is higher than its mortality rate. There are three main factors that determine the niche of a species: abiotic (environmental) conditions, biotic conditions, such as species interactions, and dispersal capacity (J. M. Soberón, 2010). These are illustrated by the BAM diagram which depicts the biotic factors (B), the abiotic factors (A), and the mobility (M) (Peterson et al., 2011).

ENM involves many steps. A very good, detailed, and recent checklist of these steps was proposed by Feng et al. (2019). For summarizing we can group them in three general steps: (1) Preprocessing, (2) Modeling, and (3) Postprocessing.

In the Preprocessing stage, acquisition and pretreatment of the species occurrence records and selecting predictor variables takes place. Databases, such as speciesLink (Canhos et al., 2015) and GBIF (Edwards, 2004), provide these records. Pretreatment is a very important and decisive step to get a good result of the proposed model. Despite the large amount of data available, it is extremely important to proceed with the application of techniques to clean and check the quality of the data, applying geographic and taxonomic filters, as described above. Abiotic variables can be downloaded as well, for example, in climatology databases such as Worldclim (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) and Bio-ORACLE (terrestrial and marine respectively) (Tyberghein et al., 2012). The environmental layers are downloaded and converted so that they can be used as input to modeling algorithms along with occurrence points. These datasets are usually in raster format, that is, a grid of two-dimensional cells, where each cell has a value. The resolution of the dataset is given by the size of a cell and the smaller the cell, the higher the resolution. In this step, we also verify the sample bias, using spatial filters—to remove points very close geographically, in order to select points with a minimum geographical distance between them (Boria, Olson, Goodman, & Anderson, 2014; Naimi, Skidmore, Groen, & Hamm, 2011; Varela, Anderson, García-Valdés, & Fernández-González, 2014) aiming to minimize the effects of the spatial autocorrelation (Dormann et al., 2007)—and applying techniques to remove cross-correlation between the predictive environmental variables, known as multicollinearity (Dormann et al., 2013). These procedures aim to reduce the spatial bias effects of the data, this spatial bias can increase the uncertainty of the models generated. Moreover, the use of clean data generates models with greater predictive power (Aiello-Lammens, Boria, Radosavljevic, Vilela, & Anderson, 2015; Calabrese, Certain, Kraan, & Dormann, 2014; Lahoz-Monfort, Guillera-Arroita, & Wintle, 2014).

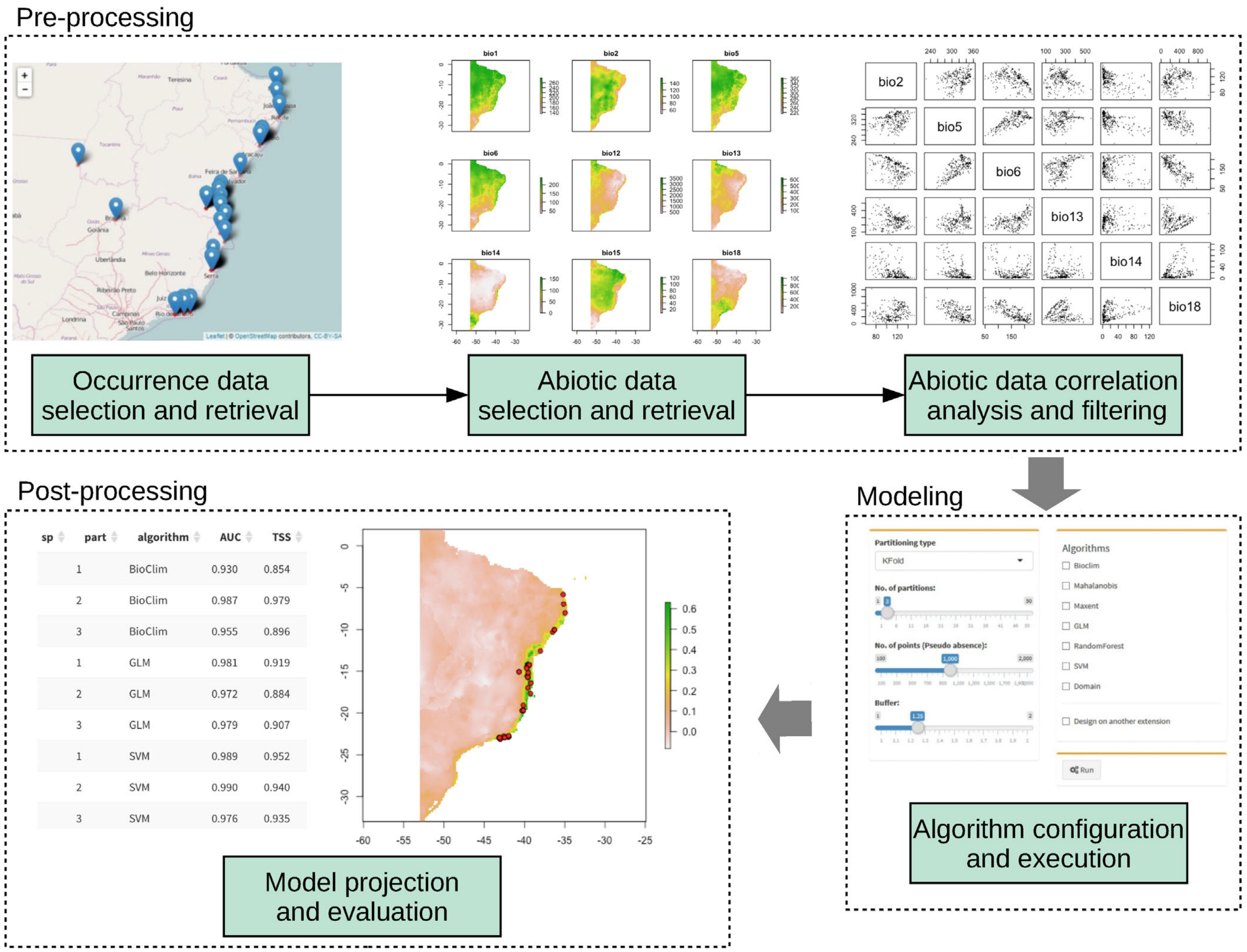
The modeling step consists of the application of algorithms to the data obtained in the preprocessing phase, for the creation of the models. Various algorithms are used in ENM, some based on machine learning, statistical inference, distance, or environmental envelopes. Machine learning algorithms include, for example, Maxent (Phillips et al., 2006) and Boosted Regression Trees (BRTs; Elith, Leathwick, & Hastie, 2008). This modeling step comprises many aspects related to the parameterization of the algorithms used, as features regularizations and learning rates. Retaining and informing the algorithm parameterization set used is very important for fine-fitting the model and for reproducibility (Qiao, Soberón, & Peterson, 2015).

Postprocessing consists of evaluating the performance of the generated model, to increase reliability, or to reduce the uncertainty of the models generated by different algorithms. The evaluation is often made from the comparison of the generated results against distribution data of the species not used in the modeling process. The indices used to measure model performance can be threshold-independent (not based on a specific threshold only), like the area under the receiver operating characteristic curve (ROC-AUC) or threshold dependent, applied in order to scale, optimize, balance, or equalize one or another type of error in the evaluation process, which is the omission or commission associated with the data set, such as Kappa (Cohen's Kappa Statistic), True Skill Statistics (TSS), all of them from the rates calculated by a confusion matrix (Allouche, Tsoar, & Kadmon, 2006; Elith et al., 2006; Pearson, Raxworthy, et al., 2006; C. Liu, White, & Newell, 2011) based on presence and absence dataset. However, the use of absence-based assessments is highly criticized, given that this is not a data usually collected and available, therefore, different methods are used to generate them, which can cause a lot of noise in the assessments performed (Barve et al., 2011; Lobo, Jiménez-Valverde, & Real, 2008; Peterson, Papeş, & Soberón, 2008). Furthermore, complex strategies with a combination of maps, considering, for example, geographic barriers, deforested areas (Anderson, Lew, & Peterson, 2003; C. Liu, Berry, Dawson, & Pearson, 2005; Pearson, Raxworthy, Nakamura, & Townsend Peterson, 2006), and multidimensional analyses (Diniz-Filho et al., 2009) can be applied. A consensus model can be generated from means of combined projection techniques, where high suitability areas coincide in most of the models generated for a given species (Araujo & New, 2007). Projection techniques in different space and time need to adopt measures that guarantee the transferability of the generated model, especially when extrapolations are expected - environmental values outside the domain of values set to fit the model (Feng et al., 2019; Owens et al., 2013).

All these steps above, illustrated in Figure 5, are fundamental for the reproducibility of the models and processes generated. So, to ensure that these processes can be reproduced, it is necessary to ensure that all this information is maintained and made available in an appropriate format.

Currently, there are several packages available in R that can help in producing useful data in the reproducibility processes of the ENM experiments (Cobos, Peterson, Barve, & Osorio-Olvera, 2019; de Andrade, Velazco, & De Marco Júnior, 2020; Golding et al., 2018; Kass et al., 2018; Qiao et al., 2016; Sánchez-Tapia et al., 2018). A framework for scalable and reproducible ENM Model-R (Sánchez-Tapia et al., 2018) was developed with the objective of unifying and automating preprocessing, processing, and postprocessing steps, as well to maintain all this information for reproducibility uses. This tool includes packages related to retrieving and cleaning data, multi-projection tools that can be applied to different temporal and spatial datasets, and postprocessing tools linked to the generated models. The entire modeling process can be parameterized using command-line tools, a local graphical user interface, or through the web.

So far, ENM has relied mostly on abiotic variables. A challenge is to incorporate biotic information as well, such as species interactions. According to (Peterson et al., 2011), these are hard to incorporate, because they are dynamic, such


Pre-processing
bio1
bio2
bio5
bio6
bio12
bio13
bio14
bio15
bio18
bio2
bio5
bio6
bio13
bio14
bio18
Occurrence data selection and retrieval
Abiotic data selection and retrieval
Abiotic data correlation analysis and filtering
Modeling
Algorithm configuration and execution
Post-processing
Model projection and evaluation
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-60
-55
-50
-45
-40
-35
-30
-25
0
-5
-10
-15
-20
-25
-30

sp	part	algorithm	AUC	TSS
	1	BioClim	0.930	0.854
	2	BioClim	0.987	0.979
	3	BioClim	0.955	0.896
	1	GLM	0.981	0.919
	2	GLM	0.972	0.884
	3	GLM	0.979	0.907
	1	SVM	0.989	0.952
	2	SVM	0.990	0.940
	3	SVM	0.976	0.935

Partitioning type: KFold

No. of partitions: 3 (1–50)

No. of points (Pseudo absence): 1,000 (100–2,000)

Buffer: 1.25 (1–2)

Algorithms

□ Bioclim

□ Mahalanobis

□ Maxent

□ GLM

□ RandomForest

□ SVM

□ Domain

□ Design on another extension

Run

FIGURE 5 Typical ENM steps comprising (1) preprocessing (occurrence data selection and retrieval, abiotic data selection and retrieval, and abiotic data correlation analysis and filtering), (2) modeling (algorithm configuration and execution), and (3) postprocessing (model projection and evaluation)

that one would need moment specific and context-specific summaries of the biotic, or interactive, variables. One experiment using biotic information was performed by (Heikkinen, Luoto, Virkkala, Pearson, & Körber, 2007) incorporating mutualism information to model four bird species, which improves the accuracy of prediction considerably.

4.2 | Biodiversity data mining

The exponential growth of data in recent years has led to an increasing number of discussions around the need for research into new methods of accessing, analyzing, and managing biological data (Howe et al., 2008). Interest in database knowledge discovery began, historically recorded, in 1989 with the *Workshop on Knowledge Discovery in Databases* (Piatetsky-Shapiro & Frawley, 1989) and has evolved greatly in recent decades. Thus, as a branch of artificial intelligence, data extraction, and knowledge discovery aims to automatically discover statistical rules and models from data. The difficulty of discovering patterns in large databases (Han, Kamber, & Pei, 2011), such as GBIF, demands other methods to access and manage biological data (Howe et al., 2008). As biodiversity databases have observed a substantial increase in data, knowledge extraction from them has become a challenge (Drew, 2011). In this sense, Hochachka et al. (2007) argues that for the development of ecological analyses where there is little prior knowledge and hypotheses are not clearly developed, exploratory analyses with data mining techniques (Liao, Chu, & Hsiao, 2012) are more appropriate than the confirmatory analyses, that is designed to test hypotheses or estimate model parameters. In ecology, some research using data mining was conducted by Spehn and Korner (2009). Pino-Mejías et al. (2010) used classification algorithms for predicting the potential habitat of species; a decision tree algorithm was also used for forest growing stock modeling (Debeljak, Poljanec, & Ženko, 2014); Kumar, Mills, Hoffman, and Hargrove (2011) used cluster analysis to identify regions with similar ecological conditions, Flügge, Olhede, and Murrell (2014) used multivariate spatial associations for grouping species into disjunct sets with similar co-association values. One of the many possibilities of using data mining was investigated by Silva (L. A. E. Silva, Siqueira, et al., 2016), which developed a methodology to allow for the application of association analysis for extracting patterns of co-occurrence from a dataset from the 50 ha Forest Dynamics Project on Barro Colorado Island, finding patterns of positive and negative correlation. To do this, association analysis was applied with the Apriori algorithm (Agrawal & Srikant, 1994). Ciarleglio, Wesley Barnes, and Sarkar (2009) proposed ConsNet, a software for designing conservation area networks using tabu search (Glover, 1986) with multi-criteria objectives. Knowledge discovery from data has been successfully used in several traditional areas such as marketing, medicine, economics, engineering, business administration; and geography (Mills, Hoffman, Kumar, & Hargrove, 2011). In Ecology, some research was also observed using data mining techniques, as in works using classification algorithms (Cutler et al., 2007; Dlamini, 2011; Hochachka et al., 2007; Lorena et al., 2011; Pino-Mejías et al., 2010); cluster analysis (Brandao et al., 2009; Kumar et al., 2011); but very little compared to other areas (Inman-Narahari, Giardina, Ostertag, Cordell, & Sack, 2010). Next, two categories of data mining algorithms that are frequently applied to Biodiversity analysis are described.

Association analysis is one of the most popular unsupervised methods of data mining for finding frequent item sets from database-logged transactions, by extracting association rules between items present in transactions, without regard to the implications of causality (Agrawal & Srikant, 1994; Han et al., 2011; Tan, Kumar, & Srivastava, 2002; Wu et al., 2008). Association analysis aims to present rules that are often unclear. One example of the application of association rules applied to the identification of species co-occurrence patterns (G. G. Z. Silva, Green, et al., 2016) The algorithms used in cluster analysis, or simply clustering, are intended to partition a set of records into groups such that records within a group are similar to each other, and records belonging to two different groups have different characteristics. The Expectation Maximization algorithm (EM), for example, uses the probability distribution to represent each cluster. It uses Gaussian probability based on density estimation theory. The algorithm makes an initial prediction for the parameters and then improves them iteratively. It generates clusters of similar size, and spherical shape, easily perceived by human eyes. Brandao et al. (2009) used EM to analyze a database of bromeliads, identifying altitudinal patterns at different spatial scales. From the results found, the use of the algorithm was recommended for the conservation of threatened species.

4.3 | Wildlife health monitoring

A comprehensive approach for wildlife health monitoring involves many challenges that need to be addressed in order to result in a globally effective mechanism for diseases prevention, such as the difficulty and limited access to wild, mostly uninhabited areas; how to overcome the high diversity and complexity of parasites, vectors, hosts and disease ecology; the methodology and infrastructure for properly collecting, storing and managing georeferenced high-quality data; how to integrate specialists from different areas to handle data, species and distinct socioenvironmental contexts; the research on knowledge extraction from data-driven models to understand, identify and predict risks to ultimately convey relevant information to society; and finally, how to sensitize decision-makers about the importance of monitoring as well as to engage the population as committed citizen scientists. The challenges are yet more acute in megadiverse countries which, in addition to biodiversity richness, usually also have to cope with vast territorial distances and sociocultural diversity.

He et al. (2016) present the eMammal framework for wildlife monitoring supported by citizen scientists. Animal images collected with camera traps are sent to its database where visual animal recognition techniques are applied. The species identification recommendations generated are reviewed by citizen scientists and, subsequently, by experts. The resulting validated records are made available to wildlife and ecological researchers. eBird (Sullivan et al., 2014) also leverages the capability of citizen scientists to gather bird observation records. Automated data quality filters are used to support species identifications performed by citizen scientists. The Brazilian Wildlife Health Information System (SISS-Geo; Chame et al., 2019), for example, is a platform for collaborative monitoring that intends to overcome the challenges in wildlife health. It aims integration and participation of various segments of society, encompassing: the registration of occurrences by citizen scientists; the reliable diagnosis of pathogens from the laboratory and expert networks; and computational and mathematical challenges in analytical and predictive systems, knowledge extraction, data integration, and visualization, and geographic information systems. It has been successfully applied to support decision-making on recent wildlife health events, such as a recent Yellow Fever epizooty (Couto-Lima et al., 2017; Moreira-Soto et al., 2018).

By automating the search for occurrence patterns, the information reaches more efficiently citizens nationwide, from the general population through experts, as well as provides the opportunity for the acquisition of knowledge about the possible patterns and parameters that contribute to the occurrence of diseases. In the medium- and long-term it also builds the capacity of researchers to develop complex modeling in the ecology of diseases that can possibly exploit geographic information in order to improve accuracy. Moreover, occurrence patterns yield data that can assist national policy on health and on biodiversity conservation.

Machine learning has been used for image analysis in wildlife monitoring, such as for automated species classification. As mentioned earlier, in (Z. He et al. 2016) the authors describe how species recognition is tackled within the eMammal cyber-infrastructure from camera-trap digital images. Once an animal (or group of) crosses the motion sensor and triggers the camera, the resulting sequence of captured images is processed in order to detect and segment the animal from the natural scene. *Detection* is the task of identifying the bounding box within the animals lie on the image, whereas *segmentation* is separating the animals (foreground) from the scene (background). Both tasks are challenging in this context—sometimes even for humans—as the animals are quite camouflaged by the heavy amount of natural elements in the wild. The approach developed to tackle this object-cutting problem (Ren, Han, & He, 2013) takes into consideration multiple image frames from the captured sequence: a standard background-foreground classifier is applied to each frame, however, the locally obtained information is fused across all frames collaboratively; this process is repeated iteratively until the refinement converges. According to the authors, this novel technique led to an improvement in the average segmentation precision of near 15% over the state-of-the-art algorithm. After the background-foreground image segmentation, an even more challenging task takes place: the animal species recognition. Here, the segmented image patches of animals are fed into previously trained machine-learning models—built based on existing labeled images—in order to classify them by species. Since these patches usually contain some elements from the background scene, that is, the segmentation is not perfect, the recognition model has to be able to cope with a good deal of noise; to make things harder, the model also needs to recognize animals at different poses. Which machine-learning algorithm is the most adequate to train such recognition models depends mainly on the amount of available training data (G. Chen, Han, He, Kays, & Forrester, 2014). In summary, conventional supervised classification algorithms are recommended for small training datasets whereas deep-learning algorithms best fit the case of an abundance of labeled data. In (G. Chen et al., 2014), using a training dataset of 14,346 images of 20 animal species, a deep convolutional neural network (DCNN) achieved an accuracy of 38%, while a Bag-of-Words model (BoW) achieved 33%. Regardless of the learning algorithm, these species recognition results are still disappointing and of limited use. From an optimistic point of view, though, it is expected that DCNN will perform better with larger training datasets as this architecture is known for its high learning capacity (G. Chen et al., 2014). For instance, in (Gomez Villa, Salazar, & Vargas, 2017) the authors report an accuracy of 88.9% on a large dataset of near 1 million images containing 26 wild animal species. Moreover, there is evidence that increasing the deepness of the neural network architecture leads to higher performance even with no additional training data (Gomez Villa et al., 2017).

Automatically identifying animal species from images seems to be the current trend in wildlife monitoring. Another great work in this vein was recently presented in (Norouzzadeh et al., 2018), where the authors, motivated by the need to eliminate the burden of manual labeling by specialists and volunteers, proposed a deep neural network (DNN) to not only identify animals, but also to count them and describe their behavior. As discussed earlier, large datasets are required in order to harness the full potential of DNNs. By using the Snapshot Serengeti (SS) dataset, which contains a total of 10.8 million classified images of 48 species to train the deep-learning model, an impressive accuracy of 93.8% was obtained. The task of recognition was divided into two stages: in the first stage, a model was trained exclusively to separate images containing at least one animal from images without animals. Then, the second DNN model was trained to take the resulting images with animals (only a quarter of the total) and perform the extraction of information, that is, identification of species, number of animals, and their characteristics. Instead of resorting to different models for each task of the second stage, the authors opted for training a single model. The reasoning behind this choice is that (i) learning related tasks simultaneously is more efficient and (ii) a single model the advantage of a reduced amount of parameters and therefore total complexity. In the article, nine DNN architectures were tested, as well as the ensemble model consisting of all of them. For the task of detecting images that contain animals, the best accuracy of 96.8% was achieved by the Very Deep Convolutional Network architecture (VGG; Simonyan & Zisserman, 2014). Regarding the species identification, the ensemble of DNNs obtained the highest score, 94.9%, followed by the Deep Residual Learning architecture (ResNet) (Z. He et al., 2016), with 93.8%. Interestingly, when considered the relaxed definition of accuracy as being *the correct answer in the top-5 guesses by the DNN model*, which would also greatly save human labor in manual labeling, these scores further improve to 99.1% and 98.8%, respectively. Counting the number of animals in the image showed to be a hard task for DNNs. Again, the ensemble of DNNs model got the best score, correctly predicting the number of animals 63.1% of the time. If allowed to approximately count the number of animals up to an error of ±1, the accuracy climbs to 84.7%. In this task, the best individual architecture was again ResNet, achieving a score of 62.8% and 83.6% for the exact and approximate accuracies, respectively. Finally, the task of describing the characteristics of the animals aimed to detect the following attributes: standing, resting, moving, eating, and whether young animals are present. Note that this is a nonexclusive multilabel classification task, since one or more animals may exhibit multiple attributes. In this task, the ensemble of DNNs model obtained 76.2% accuracy, and the second-best, ResNet, got 75.6%.

All these enthusiastic results put Deep Learning Networks as the current state-of-the-art when it comes to pattern recognition in wildlife monitoring. This translates to a significant reduction of human labor and also opens a lot of new applications in wildlife monitoring, especially for those small projects that cannot recourse to volunteers and experts for labeling. We are approaching the stage where machine-learning models are as good as humans in species identification, and they can even excel humans in that task, eventually.

Figure 6 summarizes the wildlife health monitoring workflow, from data acquisition through computational models and knowledge extraction. An observation of a wildlife animal is registered into the process either by *passive* or *active* monitoring. Camera-trap and the like are forms of passive monitoring, whereas fieldwork observation by citizen-scientists or experts is called active monitoring. Once the observation is registered, a preprocessing step may take place, such as an image detection/segmentation algorithm or a species identification model, and then the data are stored in a database. In order to ensure data quality and include missing information, the registers may be peer-reviewed and

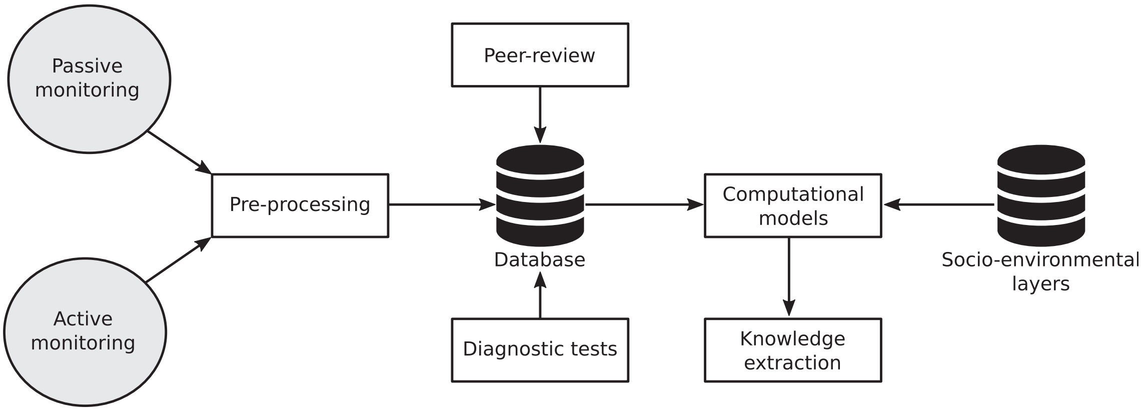


FIGURE 6 General workflow of wildlife health monitoring

edited by collaborators, where the data are expanded if necessary and verified in terms of consistency and correctness. Depending on the purpose of the monitoring, additional information may also be aggregated; in particular, it could be included *diagnostic tests* from collected animal samples, with possible indication of an ongoing epizootics. Finally, the set of acquired data, optionally coupled with extra variables (e.g., socio-environmental layers), could be used to train computational models such as *alert*, *predictive* and *forecast* models, and then be potentially used to extract knowledge about the subject being investigated (Chame et al., 2019).

4.4 | Network science applied to biodiversity

Network Science refers to a relatively new domain of scientific investigation, aiming to describe emergent properties and patterns from complex systems of interacting entities. Such relational systems are naturally represented as networks, in which interactions are represented as pairwise connections (*links*) between entities (*nodes*) and assume particular semantics depending on the nature of the modeled phenomenon. The rise of this field is strongly associated with recent advances in information technology, which provided scientists with novel tools for collecting, storing, and processing data from many knowledge domains more efficiently and in larger scales. Although a variety of networked systems in many disciplines had been studied long before that, technological advances allowed us to model real-world systems in much more detail, from large volumes data that are often public or easily accessible to the investigator. Network science (Barabási, 2016; Newman, 2010) has been applied to model networked systems in a variety of knowledge domains, including the Internet, scientific collaboration networks, and ecological networks just to cite a few (Albert & Barabási, 2002). Given their relational structure, network models are formally represented as *graphs*. Network modeling has been widely adopted in the context of biodiversity research, especially for investigating ecological and evolutionary aspects of ecosystems and natural communities. Efforts toward this goal have led to the creation of the field of *network ecology*, which has undergone a noticeable growth over the last few years (Bascompte, 2007; Borrett, Moody, & Edelmann, 2014).

Network ecology has traditionally focused on describing general aspects of the entangled networks by which organisms interact. As ecological interactions are regarded as key processes modeling ecosystems functioning and structure, unraveling their architecture and dynamics is essential for understanding a variety of ecosystem features, such as stability and energy flow. Interaction networks can be broadly classified as *food webs*, *host-parasitoid webs*, or *mutualistic webs* (Ings et al., 2009), being food webs the first ones described in literature since two classical papers by Lindeman (1942) and Odum (1956). Besides ecological interactions, network thinking has also been applied for modeling other aspects of natural systems. Patterns of animal movement can be investigated in a structured way, for instance, by means of *movement networks* (Jacoby & Freeman, 2016). These networks represent geographical space as a set of discrete and interconnected locations, forming a mesh of possible routes through which animals (or groups of animals) travel. Links between each pair of locations are weighted according to their geographical connectivity. Animal movement is thus regarded as dynamic processes composed of sequences of discrete movement steps running through the network structure. As the spatial feature is key in this type of network, they are also referred to as *spatial networks* (Bascompte, 2007).

Others have applied network science to investigate biogeographical patterns, such as species co-occurrence. C. R. Stephens et al. (2009), for instance, have used biotic interaction networks for analyzing biodiversity and predicting emerging diseases. The so-called *co-occurrence networks* model species associations in terms of their geographical distributions, such that species which are often observed occurring together in the same set of localities are considered to be strongly connected to each other. Similarly to other networked systems, co-occurrence networks are composed of a majority of the species holding co-occurrence links to very few others, while only a few species are connected to many others (M. B. Araújo, Rozenfeld, Rahbek, & Marquet, 2011). Co-occurrence network analysis has been used for many applications in biodiversity studies, such as for selecting subsets of species to be used as surrogates for the characterization of biological communities (Tulloch et al., 2016); for assessing the resilience of biotic communities toward climate change (M. B. Araújo et al., 2011); and for identifying modularity (clusters of overlapping species ranges) in biological communities from animal-location bipartite networks (Thébault, 2013).

The social network analytics framework, which is a particular application of network theory to represent and analyze social interactions in many distinct knowledge domains, has also been applied in some biodiversity studies, though in most cases for modeling animal social behavior (Faust, 2011). An alternative perspective is to look at communities of biodiversity data producers and consumers, in order to better understand the myriad of contexts in which data are collected, shared, and used. Mapping data flow within the community of biodiversity informatics initiatives, for instance, could help to prioritize and to improve the coordination of collaborative actions, leading to more effective biodiversity data-based policies (Bingham et al., 2017). Furthermore, patterns of scientific community formation can be identified and characterized by exploring collaborative paper authoring networks and scientific topic networks (Borrett et al., 2014). Analogously, a recent work (de Siracusa, Gadelha, & Ziviani, 2020) has shown that the occurrence records of species can be used to identify communities of field collectors, in terms of their mutual collaborations during fieldwork, or even in terms of their taxonomic interests. Figure 7 illustrates such networks, the data can be visualized using three perspectives: (a) Unprojected network, where collectors (green nodes) are linked to the species (red nodes) they have recorded. The total number of records of a given species by some collector is reflected in the strength of their link. (b) SCN projection onto the species set. Species are linked together if they have been collected by common collectors. The strength of links between two species is proportional to the number of collectors they share. (c) SCN projection onto the set of collectors. Collectors are linked together if they have recorded species in common. The strength of links between two collectors is proportional to the number of species they share. Link strength for both projections is graphically displayed as edges thickness. The sizes of collector and species nodes reflect their degrees, in each perspective.

As biological collections result from multiple contributions of individual collectors over time, understanding the evolution of such communities can be an invaluable surrogate for understanding the process of assembling of biological collections themselves. Such an approach opens many new perspectives of use for museum data, including characterizing sampling biases inherent to the collection. A similar example of a collaboration network in biodiversity has been presented by (Groom, O'Reilly, & Humphrey, 2014), where a correspondence network of 19th–20th century botanists was structured from digitized data from the British Herbaria. Botanists composing this network corresponded with each other by exchanging specimens, a practice that has led to the formation of exchange clubs. Many aspects regarding the particular ways botanists used to work as well as the roles they assumed could be investigated with the aid of exchange networks.

Finally, a better understanding of the factors and processes influencing the composition of species occurrence datasets would be invaluable for improving data usability, especially for species distribution modeling (Daru et al., 2017). As biological collections are typically composed of an ensemble of opportunistic species occurrence records, each of which having been gathered in a particular context by a different collection team, their datasets do not necessarily reflect the biological diversity from the areas in which the collections are physically located. Rather, they

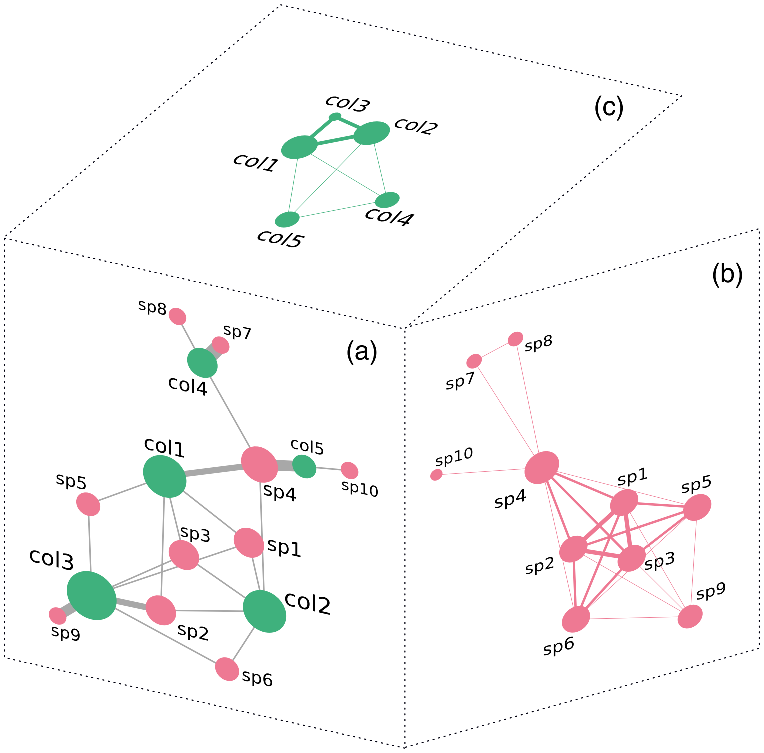


FIGURE 7 Multiple perspectives of a species-collector network (SCN)

best reflect the interests of their most active and relevant collectors, that is, those who have contributed to the collection to larger extents.

4.5 | Biodiversity genomics

The complete DNA sequence of an organism defines its genome which is present from simple species to more complex organisms such as vertebrates. With the advances in next-generation sequencing (NGS), unknown genomes have been sequenced, assembled, and deposited in public data repositories of molecular data. These data are growing fast because of the decreased cost of NGS and increased capacity of computational infrastructures (Z. D. Stephens et al., 2015; Lee & Amaro, 2018). The advances in genomic information production bring results in many application areas to society, including the production of valuable bioproducts at industrial scale (biofuels, bioenergy, cellulose fibers, gum chemicals, oils, and resins), biomonitoring of species (viruses in epidemiological surveillance [E. C. Holmes, 2008]), new drugs (vaccine design (Y. He, Preece, Hammock, Butler, & Pauw, 2015), and protein therapeutics (Leader, Baca, & Golan, 2008)).

Furthermore, molecular approaches are becoming one of the most relevant tools to support the taxonomist in species identification (Hebert, Cywinska, Ball, & DeWaard, 2003). The community is looking for the genes or regions of the genome as DNA barcode candidates, which includes cytochrome C oxidase I (COI) used to identify animals (mammals, insects, fishes); internal transcribed spacer (ITS) for fungi; ribosomal RNA (rRNA)—subunit 16S for identifying bacteria; maturase K (matK); and ribulose-bisphosphate carboxylase (rbcL) to identify plants. In this direction, Barcode of Life Data System (BOLD; Ratnasingham & Hebert, 2007) provides an integrated bioinformatics platform that assists in the acquisition, storage, analysis, and publication of DNA barcode records. It is developed and hosted by the International Barcode of Life project (iBOL), one of the largest biodiversity genomics initiatives ever executed. Hundreds of biodiversity scientists, bioinformaticians, and technologists from 25 nations are working together to construct a richly parameterized DNA barcode reference library that will be the foundation for a DNA-based identification system for all multicellular life (Page, 2008).

In general, macroscopic species are identified and cataloged by morphological aspects and receive a voucher identifier, which contains information about this species and metadata (geographical location, collector, collection date, etc.). For example, The Global Genome Biodiversity Network (GGBN) data portal (Droege et al., 2014, 2016), stores information about vouchered collections of DNA or tissue samples. Museums and institutions are joining efforts to link their biological collections with genetic data as nucleotide sequences from particular genes. Genbank (from The National Center for Biotechnology Information—NCBI), DDBJ (DNA Databank of Japan), ENA (European Nucleotide Archive), which are part of the International Nucleotide Sequence Databases Collaboration (INSDC), are the most popular repositories for nucleotide sequences. Although the voucher tag is present in Genbank since 1998, it remained poorly used (Schoch et al., 2014). Recently, the BioCollections database from NCBI connected specimen vouchers to sequence records in GenBank (Sharma et al., 2018).

The recovery of DNA of extinct species, denominated ancient DNA (or aDNA), which provides resources to understand the evolutionary process, is another promising area of biodiversity genomics. The reconstruction of aDNA involves material derived from archeological specimens, mummified tissues, preserved plant remains, and from other environments, such as permafrost and sediments (Burrell, Disotell, & Bergey, 2015). Until now, most of the extinct species sequenced belong to mammalian megafauna and ancient humans (Campbell & Hofreiter, 2012). The experimental difficulty and the challenges in this area are related to the quality of the material which is degraded with time and it is often contaminated.

Considering microscopic organisms, NGS opens a new world of capabilities reducing the need for culture isolation of microscopic species of the domains Bacteria and Archaea. This field is known as metagenomics, which is defined as the analysis of sequences taken from environmental samples, which are called metagenomes (Wooley et al., 2010). Sequencing these metagenomes produces fragments of sequences, that is, sequence reads, of organisms that are present in the environmental samples, which may belong to multiple species, and are considered extremely challenging to analyze from a computational perspective. These sequences are usually filtered to exclude those that belong to taxons that are not of interest. The resulting datasets are described using metadata standards, such as MIxS (Yilmaz et al., 2011), for supporting data discovery and mining. Next, the overlapping sequence reads are used to obtain longer sequences called *contigs* in a process known as assembly. SPAdes (Bankevich et al., 2012) is one example of a tool used for assembly. There are several important biological discoveries based on complete and near-complete genomes assembled from metagenomes. For instance, the possible ancestor of mitochondria and the possible ancestor of the first eukaryotic cells were proposed from reconstructed phylogenies from those genomes (Eme, Spang, Lombard, Stairs, & Ettema, 2017; Martijn, Vosseberg, Guy, Offre, & Ettema, 2018; Zaremba-Niedzwiedzka et al., 2017). After assembly, the resulting sequences can be analyzed for identifying genes using either sequence alignment, for genes with homologs present in public databases, or through ab initio gene prediction using, for instance, hidden Markov models. Other types of analyses involving metagenomes include the evaluation of species diversity and functional annotation (Wooley et al., 2010). Various tools compose these different analyses into metagenomic workflows. An example is MG-RAST (Meyer et al., 2008; Wilke et al., 2016), a web portal that provides metagenomic dataset analysis workflows containing activities such as quality control, similarity-based annotation, and functional and taxonomic profiling. SUPER-FOCUS (G. G. Z. Silva, Green, et al., 2016) also produces functional and taxonomic profiles from metagenomic datasets. However, its organism identification is based on alignment-free techniques used by the FOCUS (G. G. Z. Silva, Cuevas, Dutilh, & Edwards, 2014) tool. Metagenomics can support various environmental studies such as the analysis of coral diseases. Garcia et al. (2013) identified taxonomic groups that were more abundant in *Mussismilia braziliensis* corals affected by the white plague disease when compared to healthy corals of the same species. Integrating data from metagenomics with data from other aspects of biodiversity, such as species populations and environmental monitoring, is still a challenging task. More recently, there were efforts to integrate these standards (O'Tuama et al., 2012). Ongoing efforts for improving data integration in bioinformatics and biodiversity are using semantic web techniques (Walls et al., 2014). These efforts are essential in supporting integrative ecosystem studies, such as (Meirelles et al., 2015), where different attributes of the ecosystem found in the mesophotic reefs of the Vitória-Trindade seamount chain were correlated to infer its properties.

As the computational challenges in this area, we can point questions related to storage, recovery, and integration of the information; conceptual modeling, ontology, and semantic representation of the molecular domain. Furthermore, there are usually multiple computational activities in bioinformatics analyses including filtering, normalization, and annotation. Efforts to ensure reproducibility (Cohen-Boulakia et al., 2017) of these analyses involve (but are not limited to) task composition tools (scripts (Babuji et al., 2019), pipelines, scientific workflows (Liew et al., 2016), and software containers (Boettiger, 2015), web-based software platforms, such as Galaxy (Bedoya-Reina et al., 2013), commonly used applications, and source code available in repositories such as Github. We explore these issues in more detail in Section 4.6.

4.6 | Biodiversity workflows and reproducibility

Scientific data are being produced at an exponential growth rate by increasingly available scientific sensors. This, coupled with sophisticated computational models that process these data, has demanded new techniques (Hey, Tansley, & Tolle, 2009) for managing computational scientific experiments in a scalable and reproducible way. Wilson et al. (2014) propose best practices for managing scientific computations. These include: recording datasets, programs, libraries, and parameters used, including their respective identifiers or versions, to enable better reproducibility; and using high-level languages for programming and moving to lower-level languages only when performance improvement is necessary. These experiments are often specified as scientific workflows (Deelman, Gannon, Shields, & Taylor, 2009; Liew et al., 2016; Shade & Teal, 2015), which are given by a composition of computational tasks that exchange data through production and consumption relationships. A scientific workflow management system (SWMS) provides features such as fault-tolerance, scalable execution, scalable data management, data dependency tracking, and provenance recording, that greatly reduce the complexity of managing the life-cycle of these experiments (Mattoso et al., 2010). Scientific workflows are often provided through research data portals, Chard et al. (2018) present a design pattern for such portals for data-intensive scientific problems. Reproducibility (Peng, 2011) is an essential property in science. In computational research, it can be a challenging task since one might need vasts amounts of data or supercomputing resources to reproduce a result. However, the reproducibility of experiments allows the verification and validation of results by others and may increase the chances that it can be reused. This is especially relevant given the demand from journals in different domains for submissions of reproducible computational research and also because of recent initiatives that encourage greater accessibility and transparency in scientific research (Stodden, Guo, & Ma, 2013; Vicente-Saez & Martinez-Fuentes, 2018). Sandve, Nekrutenko, Taylor, and Hovig (2013) propose rules that can be followed to better support reproducibility, including recording the steps that were executed to obtain a result, archiving programs that were used in a computational experiment, and versioning the scripts and workflows used.

Meng et al. (2015) propose a framework that tackles reproducibility by providing features for sandboxing and preserving computational environments. A combination of containers (Boettiger, 2015; Hale, Li, Richardson, & Wells, 2017) and tools for intercepting system calls is used in order to achieve preservation. Many computational experiments have a detailed record of their execution, such as the datasets used and computational tasks used, and enable easier verification of results. These records describe the *provenance* (Carata et al., 2014; Freire, Koop, Santos, & Silva, 2008) of the computational experiment. It can support the reproducibility and validation of e-Science experiments. Miles et al. (2007), for instance, propose an architecture for validation of e-Science experiments based on both provenance assertions and ontologies. DataONE, for instance, included support for tracking the provenance of their datasets (Cao et al., 2016). To improve the reuse of research data, Wilkinson et al. (2016) propose a set of guidelines for scientific data or digital assets to be findable, accessible, interoperable and reusable, also known as the FAIR principles. The idea behind these principles is that the process of data recovery may be more automatic, with minimal user intervention, since many of the experiments in different domains rely on computational support for the manipulation and analysis of data. One of the motivations to follow these principles is that good data management improves the quality of the publications. The technologies and tools to achieve each of the principles can vary. In biodiversity, for example, different commonly used tools can be combined to allow experiments to be FAIR (Harjes, Link, Weibulat, Triebel, & Rambold, 2020).

Biodiversity follows the same trend of rapidly increasing production of data found in other areas of science. Currently, biodiversity data are being integrated at a global scale through initiatives such as GBIF (Edwards, 2000). Techniques for analysis and synthesis of biodiversity data, such as ENM (Elith & Leathwick, 2009; Peterson et al., 2011), are widely used. These analyses typically employ several different applications executed in a loosely coupled manner, being a typical use case for scientific workflow management tools (J. Liu, Pacitti, Valduriez, & Mattoso, 2015). Next, we list some works related to scientific workflows and reproducibility in biodiversity. Pennington et al. (2007) describe the implementation of species distribution modeling (SDM) scientific workflows using Kepler (Ludäscher et al., 2006). Their approach allows for easy management of structural aspects of the scientific workflow, such as easily replacing application components. They also developed application components for data transformation and preprocessing, geospatial processing, and semantic annotation of processes. These experiments use occurrence data from the Mammal Networked Information System (MaNIS)[12] and future climatological scenarios from IPCC to predict the climate-change impact on more than 2,000 species. Morisette et al. (2013) present the Software for Assisted Habitat Modeling (SAHM) that allows for managing the various steps of SDM, including pre- and postprocessing activities. The implementation is coupled with the Vistrails (Freire et al., 2006) scientific workflow management system, which supports provenance management. Talbert, Talbert, Morisette, and Koop (2013) also describe SAHM and analyze the data management challenges of SDM using scientific workflows. Amaral et al. (2015) present the *EUBrazilOpenBio Hybrid Data Infrastructure* which implements cloud services for the biodiversity domain such as taxonomic mapping and resolution and SDM. Scientific workflows are supported both with DAGMan (Couvares, Kosar, Roy, Weber, & Wenger, 2007) and EasyGrid AMS (Boeres & Rebello, 2004). They evaluate the execution of SDM on cloud computing resources showing good performance. Candela, Castelli, Coro, Pagano, and Sinibaldi (2016) give a detailed description of an integrated cloud-based environment for SDM of the EUBrazilOpenBio Hybrid Data Infrastructure which includes components for retrieving species occurrences, environmental layers, and execution of various models for predicting species distributions. Some SDM applications and workflows are available through web portals, such as the Biodiversity Virtual e-Laboratory (BioVel) (A. R. Hardisty et al., 2016). BioVel (A. R. Hardisty et al., 2016) offers a web-based environment for managing scientific workflows for biodiversity. Various predefined activities are available in its interface: geographical and temporal selection of occurrences (BioSTIF), data cleaning, taxonomic name resolution, ENM algorithms, population modeling, ecosystem modeling, and metagenomics and phylogenetics applications (Vicario, Balech, Donvito, Notarangelo, & Pesole, 2012).

iPlant (Goff et al., 2011) is a computational research infrastructure, or cyberinfrastructure, for plant science. Its applications include the *Tree of Life* to produce phylogenetic trees of all green plant species, and *Genotype to Phenotype* to predict plant phenotypes from their genetic data. Kurator (Dou et al., 2012) is a software package for the Kepler (Ludäscher et al., 2006) scientific workflow management system that supports composing various data curation activities into scientific workflows. Prebuilt activities include georeferencing, scientific name, and flowering time validators. Provenance is recorded to document all the transformations activities that data went through caused by the various data curation activities. Nguyen et al. (2017) developed scientific workflows for assessing ecosystem risk based on IUCN guidelines (Keith et al., 2013) that use five rule-based criteria to assign one of eight risk categories that range from *least concern* (LC) to *collapsed* (CO). The assessment is performed in two phases. First, a stochastic ecosystem model is executed for the Meso-American Reef Ecosystem risk assessment by predicting future reef properties under diverse scenarios. This step was implemented both in Nimrod/G (Abramson, Giddy, & Kotler, 2000) and Spark (Zaharia et al., 2016), for comparative purposes. The Spark version had a considerably better performance in terms of computing time. Next, a workflow was implemented in the Kepler scientific workflow management system (Ludäscher et al., 2006) to execute the IUCN ecosystem risk assessment methodology using the results of the stochastic ecosystem model execution and applying its five rule-based criteria. Reproducibility is an important property in this process since risk assessment is often re-executed and its results need to be discussed by experts and decision-makers (Guru et al., 2016).

Borregaard and Hart (2016) discuss the importance of allowing reproducibility in ecology experiments, especially due to the change in how they have been specified. Currently, scripting languages such as R and Python have been increasingly adopted in the data analysis process. In this context, one of the challenges is to provide a means for users who do not have programming language skills to replicate the experiments. Golding et al. (2018) present Zoon R, an R package that allows SDM to be reproducible and shareable through the specification of a workflow where the result is an R object that contains the data, code, and results used in the analysis. The resulting object can be published to a data repository so that others can access it, and it can be loaded back into the R environment together with the package allowing for reproducibility of the analysis. Cohen-Boulakia et al. (2017) explore the use of scientific workflows for the reproducibility of computational experiments in the life sciences. They analyze scientific workflow techniques and systems and evaluate to what extent they support reproducibility requirements in life science applications. Plant phenotyping, which evaluates how plants respond to different environmental conditions by monitoring their traits, was one of the use cases. For instance, keeping track of several different tool versions used in a workflow and their respective compatibility is one of its reproducibility requirements. They define different levels of reproducibility in the workflow context. Considering two scientific workflows A and B and assuming A has already been executed. When B is executed: *repeatability* is achieved when B contains exactly the same components of A; *replicability* is obtained when B uses similar (Starlinger, Brancotte, Cohen-Boulakia, & Leser, 2014) input components of A and both executions reach the same conclusion; *reproducibility* happens when both executions lead to the same scientific conclusion; *reusability* is observed when the specification B contains the specification of A. The authors analyze three workflow aspects from the reproducibility perspective: workflow specification, workflow execution, and workflow context, and runtime environment. Workflow specifications can support better reusability through common specification languages, such as CWL[13] (Common Workflow Language), and annotations. Assessing workflows similarity is critical for reuse but progress is still needed in solving this problem. Recording and analyzing workflow execution details can be supported by provenance information. Freire and Chirigati (2018) discuss other reproducibility levels that can be achieved and how they relate to provenance data that include aspects on the platform, implementation, and data used by an experiment. Depending on the type of provenance collected for each of these aspects, the experiment may be repeatable, re-runnable, portable, extendable or modifiable. While most systems support the PROV (Moreau, Groth, Cheney, Lebo, & Miles, 2015) standard, visualizing and analyzing large provenance datasets is still challenging. Also, preserving the runtime environment is still a challenge that is being addressed with virtualization technologies (Hale et al., 2017). WholeTale (Brinckman et al., 2019), for instance, is a computational environment that has reproducibility features. It has components for data collection, identity management, data publication, and interfaces to analytical tools, called frontends. These frontends will manipulate data and can be given, for instance, by interactive notebooks such as Jupyter.[14] The system is integrated with DataONE (Michener et al., 2012), users can search and retrieve datasets from it. Frontends are packaged as Docker containers (Boettiger, 2015) that can be executed on high-performance computing resources. The interaction between the datasets and analytical tools is documented and recorded in a metadata management system. This allows for reproducing the entire computational research performed, from data retrieval to data analysis and its outputs, including the computational environments used. Feng et al. (2019) present a checklist for maximizing the reproducibility of ENM, describing in detail each step of the process, and what should be preserved in each of them to enable better reproducibility. Mondelli, Townsend Peterson, and Gadelha (2019) presented a conceptual model and framework for supporting reproducibility and FAIR principles in computational experiments. The framework is evaluated with an ENM case study.

5 | BIODIVERSITY INFORMATICS CHALLENGES AND CONCLUDING REMARKS

The acceleration of global changes requires a constant assessment of their impacts on biodiversity and, consequently, on ecosystem services that are essential to humans. Some areas of the globe, for example, the South Atlantic Ocean, remain highly understudied, and therefore their biodiversity underestimated. A better understanding of marine biodiversity could be achieved with help of biodiversity informatics to leverage surveys to uncover novel species and systems, such as the Great Amazon Reef (Francini-Filho et al., 2018). To address this problem, biodiversity data must be systematically collected and analyzed. In this context, biodiversity informatics is an essential collection of methodologies, tools, and techniques to achieve this goal. EBVs (Pereira et al., 2013) were proposed as a set of indicators that would allow for systematic monitoring of biodiversity. However, the production of these indicators is still a challenge (Peterson & Soberón, 2017), in particular regarding the existence of information gaps that can prevent global-scale inferences on the state of biodiversity. These inferences provide essential input to decision-makers in devising governmental policies toward meeting global targets on biodiversity conservation, such as the Aichi Biodiversity Targets. The Bari Manifesto (A. R. Hardisty et al., 2019) was proposed as a set of guidelines for biodiversity informatics infrastructures to enable the implementation of scientific workflows for measuring or estimating EBVs gathering data from potentially multiple infrastructures and countries. In this section, we describe existing challenges for biodiversity informatics to become a systematic and global-scale tool for monitoring and making inferences about biodiversity. In Table 4, we summarize various tools and databases surveyed along with which steps of the biodiversity informatics life cycle they approach.

As described in Section 3.2, the availability of detailed information about most organisms is still very scarce (Peterson, 2006). This hinders the usage of these data in many biodiversity data analysis applications, such as ENM. Furthermore, we described other issues with biodiversity data, such as biases and frequent taxonomic and georeferencing errors. Therefore, one of the challenges of biodiversity informatics is not only to increase the amount of available data, filling some of the existing gaps, but also to reduce its bias and improve its quality. Some promising work addressing these issues are listed next:

- Heidorn (Heidorn, 2008) observes that data from smaller scientific projects are rarely available to other scientists even though their aggregated size and value for research are considerable. This phenomenon is denominated the *long tail of science*. In biodiversity and ecology, some progress has been achieved through projects such as GBIF and DataONE, that receive a considerable amount of their datasets from small research groups. One issue with making these datasets available is the effort required to map the concepts present in them to standard vocabulary terms used in major biodiversity databases. Entity resolution techniques (Köpcke, Thor, & Rahm, 2010) have the potential to assist and speed up these record linkage routines.
- Data collection could be substantially intensified by applying artificial intelligence methods for automating specimen identification, some preliminary work in this direction include the application of deep learning techniques for species identification in herbarium sheets (Carranza-Rojas et al., 2017; Carranza-Rojas, Joly, Goëau, Mata-Montero, & Bonnet, 2018).
- Remote sensing provides the opportunity to observe the Earth regularly and, therefore, could benefit biodiversity monitoring by increasing the amount of data collected. It can also be a valuable tool to observe areas that are difficult to access through field expeditions. One of the pioneering works in this area was proposed by Holden and Ledrew (1999) by using hyperspectral remote sensing to monitor coral reefs. Clark et al. (2005) identified tree species using remote sensing images. Fretwell et al. (2012) were able to use satellite-based remote sensing to survey the Emperor Penguin on a global scale. Fernández et al. (2020) have suggested the use of remote sensing and on-site observations for estimating EBVs using biodiversity modeling.
- As observed in Section 3.2, assessing the quality of a dataset is a critical step for any subsequent analysis and synthesis activity that might use it. Users should establish the intended use of datasets in their research. Determining if a dataset is fit for a particular use is still a challenge in biodiversity informatics since records available in public databases contain various types of errors. A promising approach was proposed by Veiga et al. (2017) composed of a framework for biodiversity data quality assessment and management that allows for users to define their data quality requirements and when a particular dataset is fit-for-use in a standardized manner. P. J. Morris et al. (2018) made some progress by implementing a library of small data quality assessment routines that can be composed into more complex workflows, to report data quality in terms of the framework proposed by Veiga et al. (2017).

In the GBIO (Hobern et al., 2013) report, produced by leading biodiversity informatics researchers, a number of areas of biodiversity informatics of limited or minimal progress were identified and can be considered research challenges. Biological systems modeling was considered an area of research in biodiversity informatics with minimal progress. Advances in this area could be composed of computational, or *in-silico* models or simulations ranging from single organisms to entire ecosystems. Current temporal and spatial modeling in biodiversity, such as ENM, described in

TABLE 4 A selection of biodiversity informatics databases and tools classified according to target life-cycle step: data planning and collection (DC), data quality and fitness-for-use (DQ), data description (DD), data preservation and publication (DP), data discovery and integration (DI), and computational modeling and data analysis (CM)

Tool or database name	Reference	DC	DQ	DD	DP	DI	CM
DMPTool	(Strasser et al., 2014)	✓					
Data stewardship wizard	(Pergl et al., 2019)	✓					
Morpho	(Higgins et al., 2002)	✓	✓	✓	✓		
Metacat	(Berkley et al., 2001)	✓	✓	✓	✓	✓	
Catalog of life	(Roskov et al., 2013)	✓	✓	✓	✓	✓	
BHL	(Gwinn & Rinaldo, 2009)	✓	✓	✓	✓	✓	
eBird	(Sullivan et al., 2014)	✓	✓	✓	✓	✓	
eMammal	(Z. He et al., 2016)	✓	✓	✓	✓	✓	
Brazilian Flora 2020	(Forzza et al., 2012)	✓	✓	✓	✓	✓	
BRAHMS	(Filer, 2013)	✓	✓	✓	✓	✓	
Jabot	(da Silva et al., 2017)	✓	✓	✓	✓	✓	
MorphoBank	(O'Leary & Kaufman, 2011)	✓	✓	✓	✓	✓	
BEXIS 2	(Gerlach et al., 2015)	✓	✓	✓	✓	✓	
Atlas of living Australia	(Belbin & Williams, 2016)	✓	✓	✓	✓	✓	✓
speciesLink	(Canhos et al., 2015)	✓	✓	✓	✓	✓	✓
GenBank	(Benson et al., 2013)	✓	✓	✓	✓	✓	✓
MG-RAST	(Meyer et al., 2008)	✓	✓	✓	✓	✓	✓
BOLD	(Ratnasingham & Hebert, 2007)	✓	✓	✓	✓	✓	✓
iPant	(Goff et al., 2011)	✓	✓	✓	✓	✓	✓
VoSeq	(Peña & Malm, 2012)	✓	✓	✓	✓	✓	✓
SISS-geo	(Chame et al., 2019)	✓	✓	✓	✓	✓	✓
BioGeomancer	(R. P. Guralnick et al., 2006)		✓				
Taxamatch	(Rees, 2014)		✓				
Geospatial data quality	(Otegui & Guralnick, 2016)		✓				
TNRS	(Boyle et al., 2013)		✓				
Kurator	(P. J. Morris et al., 2018)		✓				
BioVel	(A. R. Hardisty et al., 2016)		✓			✓	✓
EU-Brazil OpenBio	(Amaral et al., 2015)		✓			✓	✓
Model-R	(Sánchez-Tapia et al., 2018)		✓			✓	✓
GGBN	(Droege et al., 2014)		✓	✓		✓	
BioCollections	(Sharma et al., 2018)		✓	✓		✓	
GBIF	(Edwards, 2004)		✓	✓	✓	✓	
DataONE	(Michener et al., 2012)		✓	✓	✓	✓	
OBIS	(Grassle, 2000)		✓	✓	✓	✓	
SiBBr	(Baringo Fonseca, Correa, Soto, & Sacramento, 2017)		✓	✓	✓	✓	
IPT	(Robertson et al., 2014)			✓	✓		
Scratchpads	(Smith et al., 2011)				✓	✓	✓
WholeTale	(Brinckman et al., 2019)					✓	✓
Maxent	(Phillips et al., 2006)						✓
OpenModeller	(Souza Muñoz et al., 2009)						✓
Kuenm	(Cobos et al., 2019)						✓

(Continues)

TABLE 4 (Continued)

Tool or database name	Reference	DC	DQ	DD	DP	DI	CM
NicheA	(Qiao et al., 2016)						✓
ZOON	(Golding et al., 2018)						✓
Wallace	(Kass et al., 2018)						✓
ENKTML	(de Andrade et al., 2020)						✓
SAHM	(Morisette et al., 2013)						✓
SUPER-FOCUS	(G. G. Z. Silva, Green, et al., 2016)						✓

Section 4.1, rely on species occurrence data. More fine-grained modeling would require incorporating species trait data (Schneider et al., 2019). Cardinale et al. (2012), for instance, advocated the development of new predictive models that take into account species interactions to predict the impact of biodiversity on ecosystem processes based on species traits. Areas of limited progress identified by the GBIO included:

- Automated remote-sensed observation has the potential to enable observation of biodiversity in large and remote areas. More recently, preliminary work was conducted on defining biodiversity indicators that could be derived from images collected by satellite remote sensing (Pettorelli et al., 2016) and processed using statistical analysis and classification algorithms.
- Identifying trends and making predictions about biodiversity could determine future trends in biodiversity under different global change scenarios. Ongoing research in this area includes predicting how climate change will affect species distributions (de Siqueira & Peterson, 2003; Thomas et al., 2004; Pearson et al., 2006; M. Araújo et al., 2008; Wiens et al., 2009; M. B. Araújo & Peterson, 2012) and zoonotic diseases (Estrada-Peña, Ostfeld, Peterson, Poulin, & de la Fuente, 2014).
- Providing access to aggregate species trait data, consisting of data on species characteristics and their interactions. Data are incomplete, there are not much data about relative abundances of species, their traits, and on how they interact. This information is needed for creating better models to study ecosystem processes. This type of data can enable more complex biological systems modeling, such as evolutionary inference. MorphoBank (O'Leary & Kaufman, 2011), for instance, allows for scientists to upload images with the morphology of organisms with associated data.

Scientists often need to combine multiple sources of data (Fujioka et al., 2014; Jones et al., 2006) in biodiversity analysis and synthesis activities. Although there are many gaps in biodiversity data, such as the reduced availability of species trait data, there are many machine-readable and freely available, that is, open, datasets (Reichman et al., 2011) from areas such as remote sensing, socioeconomics, and climatology, that can be integrated into biodiversity studies. Open data are widely available online, including data provided by many governments. However, it is highly heterogeneous, dispersed in multiple sources, and may not provide metadata or schema. Metadata, described in Section 3.3, is helpful in discovering datasets and in integrating them when dataset attribute definitions are provided, as it is possible with EML (Fegraus et al., 2005). Semantic web (Walls et al., 2014), as described in Section 3.5, can also support data integration through the use of various existing ontologies for biodiversity and other domains. However, their increased usefulness depends on the widespread adoption of ontologies and metadata standards by data providers, a process that is still underway. A promising approach to overcome these limitations has been to use machine learning techniques to support open data integration activities (Dong & Rekatsinas, 2018; Miller, 2018), such as entity matching (Mudgal et al., 2018; Nargesian, Zhu, Pu, & Miller, 2018). These recently proposed techniques could be leveraged and extended for integrating biodiversity and other related datasets.

In Table 5, we describe computational techniques used in each step of the biodiversity informatics life cycle that were be explored in this work.

Trends, indicators, and facts derived from biodiversity data analysis and synthesis activities might be used for guiding governmental decision making in critical areas such as conservation area planning, impact assessment of large construction work projects, and zoonotic disease prevention. Since such decisions can have large-scale impacts in society, being able to trace back the processes involved in reaching them is an essential property. Therefore, it is important to

TABLE 5 Computational techniques used in the biodiversity informatics life cycle: string processing (SP), metadata management (MD), conceptual modeling (CM), semantic web (SW), machine learning (ML), statistics (ST), geographical information systems (GS), graph theory (GT)

Life cycle step	SP	MD	CM	SW	ML	ST	GS	GT
Data planning and collection					✓			
Data quality	✓				✓		✓	
Data description		✓						
Data publication			✓					
Data discovery and integration				✓	✓			
Ecological niche modeling					✓	✓	✓	
Wildlife health analysis					✓	✓	✓	
Biodiversity data mining					✓	✓	✓	
Biodiversity networks								✓
Biodiversity genomics	✓				✓	✓		✓

use methodologies and techniques that are reproducible (Cohen-Boulakia et al., 2017; Ivie & Thain, 2018; Peng, 2011; Sandve et al., 2013) when executing these activities. Some initial advances were achieved in projects such as DataONE (Cao et al., 2016) and WholeTale (Brinckman et al., 2019) by recording the provenance of biodiversity datasets and of their analysis and synthesis. However, reproducibility in computational science, in general, is still a challenge. Providing reproducible frameworks for biodiversity analysis and synthesis activities would enable better decision traceability and validation of trends and indicators produced by them.

ACKNOWLEDGMENTS

The work is partially supported by CAPES (grant numbers 441929/2016-8 and 461572/2014-1), CNPq (grant numbers 441929/2016-8, 461572/2014-1, and 381427/2017-0), Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2051 – Project-ID 390713860, and FAPERJ. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Luiz Gadelha: Conceptualization; investigation; writing-original draft; writing-review and editing. **Pedro de Siracusa:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Eduardo Dalcin:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Luís da Silva:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Douglas Augusto:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Eduardo Krempser:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Helen Affe:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Raquel Costa:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Maria Luiza Mondelli:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Pedro Meirelles:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Fabiano Thompson:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Marcia Chame:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Artur Ziviani:** Conceptualization; investigation; writing-original draft; writing-review and editing. **Marinez de Siqueira:** Conceptualization; investigation; writing-original draft; writing-review and editing.

ORCID

Luiz M. R. Gadelha Jr https://orcid.org/0000-0002-8122-9522

ENDNOTES

[1] http://www.catalogueoflife.org

[2] http://www.marinespecies.org/

[3] https://www.geo-locate.org

[4] http://specifysoftware.org

[5] https://emu.axiell.com

[6] http://www.bg-base.com

[7] http://dublincore.org/

[8] http://www.gbif.org/dataset/1edcfe6d-da55-4d59-b30e-468cd21f8b0b

[9] https://www.gbif.org/publisher/4205110f-3f0f-40d8-bd0f-2fa71bc827b5

[10] https://cran.r-project.org/web/packages/rgbif/

[11] https://recology.info/2015/11/pygbif/

[12] http://manisnet.org

[13] http://www.commonwl.org/

[14] http://jupyter.org

RELATED WIREs ARTICLES

Tutorial on biological networks

FURTHER READING

Thompson, F. L. (2013). Metagenomic analysis of healthy and white plague-affected *Mussismilia braziliensis* corals. *Microbial Ecology*, *65*(4), 1076–1086. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23314124. https://doi.org/10.1007/s00248-012-0161-4

REFERENCES

Abramson, D., Giddy, J., & Kotler, L. (2000). *High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid?* Proceedings 14th International Parallel and Distributed Processing Symposium. Cancun, Mexico: IPDPS 2000, IEEE Computer Society. pp. 520–528. Retrieved from http://ieeexplore.ieee.org/document/846030/ https://doi.org/10.1109/IPDPS.2000.846030

Agrawal, R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*. Proceedings of the 20th VLDB Conference. Santiago de Chile, Chile, pp. 487–499.

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*(5), 541–545. Retrieved from http://doi.wiley.com/10.1111/ecog.01132. https://doi.org/10.1111/ecog.01132

Ailamaki, A., Kantere, V., & Dash, D. (2010, jun). Managing scientific data. *Communications of the ACM*, *53*(6), 68. Retrieved from http://dl.acm.org/ftgateway.cfm?id=1743568&type=html–78. https://doi.org/10.1145/1743546.1743568

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. Retrieved from https://link.aps.org/doi/10.1103/RevModPhys.74.47. https://doi.org/10.1103/RevModPhys.74

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. Retrieved from http://doi.wiley.com/10.1111/j.1365-2664.2006.01214.x. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Amaral, R., Badia, R. M., Blanquer, I., Braga-Neto, R., Candela, L., Castelli, D., … Torres, E. (2015). Supporting biodiversity studies with the EUBrazilOpenBio hybrid data infrastructure. *Concurrency and Computation: Practice and Experience*, *27*(2), 376–394. Retrieved from http://doi.wiley.com/10.1002/cpe.3238. https://doi.org/10.1002/cpe.3238

Anderson, R. P., Lew, D., & Peterson, A. (2003). Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecological Modelling*, *162*(3), 211–232. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304380002003496. https://doi.org/10.1016/S0304-3800(02)00349-6

Araujo, M., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, *22*(1), 42–47. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S016953470600303X. https://doi.org/10.1016/j.tree.2006.09.010

Araújo, M., Nogués-Bravo, D., Reginster, I., Rounsevell, M., & Whittaker, R. (2008). Exposure of European biodiversity to changes in human-induced pressures. *Environmental Science & Policy*, *11*(1), 38–45. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1462901107000780. https://doi.org/10.1016/j.envsci.2007.07.002

Araújo, M. B., & Peterson, A. T. (2012). Uses and misuses of bioclimatic envelope modeling. *Ecology*, *93*(7), 1527–1539. https://doi.org/10.1890/11-1930.1

Araújo, M. B., Rozenfeld, A., Rahbek, C., & Marquet, P. A. (2011). Using species co-occurrence networks to assess the impacts of climate change. *Ecography*, *34*(6), 897–908. https://doi.org/10.1111/j.1600-0587.2011.06919.x

Araújo, M. B., & Williams, P. H. (2000). Selecting areas for species persistence using occurrence data. *Biological Conservation*, *96*(3), 331–345. https://doi.org/10.1016/S0006-3207(00)00074-4

Babuji, Y., Woodard, A., Li, Z., Katz, D. S., Clifford, B., Kumar, R., … Chard, K. (2019). *Parsl: Pervasive Parallel Programming in Python*. 28th ACM International Symposium on High-Performance Parallel And Distributed Computing (HPDC). Phoenix, Arizona. https://doi.org/10.1145/3307681.3325400

Balmford, A., Bennun, L., Brink, B. T., Cooper, D., Côte, I. M., Crane, P., … Walther, B. A. (2005). Ecology: The convention on biological Diversity's 2010 target. *Science (New York, N.Y.)*, *307*(5707), 212–213. Retrieved from http://www.sciencemag.org/content/307/5707/212.short. https://doi.org/10.1126/science.1106281

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barabási, A.-L. (2016). *Network science*, Cambridge, England: Cambridge University Press.

Baringo Fonseca, C., Correa, L., Soto, N., & Sacramento, R. (2017). SiBBr: Envisioning the spatial distribution of Brazilian biodiversity records. *Proceedings of TDWG*, *1*, e19966. Retrieved from https://tdwgproceedings.pensoft.net/articles.php?id=19966. https://doi.org/10.3897/tdwgproceedings.1.19966

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., … Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222*(11), 1810–1819. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0304380011000780. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Bascompte, J. (2007). Networks in ecology. *Basic and Applied Ecology*, *8*(6), 485–490. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1439179107000576. https://doi.org/10.1016/j.baae.2007.06.003

Bedoya-Reina, O. C., Ratan, A., Burhans, R., Kim, H. L., Giardine, B., Riemer, C., … Miller, W. (2013). Galaxy tools to study genome diversity. *GigaScience*, *2*(1), 17. Retrieved from http://www.gigasciencejournal.com/content/2/1/17. https://doi.org/10.1186/2047-217X-2-17

Belbin, L., & Williams, K. J. (2016). Towards a national bio-environmental data facility: Experiences from the atlas of living Australia. *International Journal of Geographical Information Science*, *30*(1), 108–125. Retrieved from http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1077962. https://doi.org/10.1080/13658816.2015.1077962

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic acids research*, *41*(Database issue), D36–D42. Retrieved from http://nar.oxfordjournals.org/content/41/D1/D36. https://doi.org/10.1093/nar/gks1195

Berendsohn, W. G. (1995). The concept of "potential taxa" in databases. *Taxon*, *44*(2), 207–212. https://doi.org/10.2307/1222443

Berendsohn, W. G. (1997). A taxonomic information model for botanical databases: The IOPI model. *Taxon*, *46*(2), 283–309. https://doi.org/10.2307/1224098

Berendsohn, W. G., Güntsch, A., Hoffmann, N., Kohlbecker, A., Luther, K., & Müller, A. (2011). Biodiversity information platforms: From standards to interoperability. *ZooKeys*, *150*, 71–87. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3234432&tool=pmcentrez&rendertype=abstract. https://doi.org/10.3897/zookeys.150.2166

Berkley, C., Jones, M., Bojilova, J., & Higgins, D. (2001). *Metacat: A Schema-independent XML Database System*. Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001, Fairfax, Virginia: IEEE Computer Society. pp. 171–179. Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=938549 https://doi.org/10.1109/SSDM.2001.938549

Bingham, H., Doudin, M., Weatherdon, L., Despot-Belmonte, K., Wetzel, F., Groom, Q., … Martin, C. (2017). The biodiversity informatics landscape: Elements, connections and opportunities. *Research Ideas and Outcomes*, *3*, e14059. Retrieved from http://riojournal.com/articles.php?id=14059. https://doi.org/10.3897/rio.3.e14059

Bisby, F. A. (2000). The quiet revolution: Biodiversity informatics and the internet. *Science*, *289*(5488), 2309–2312. Retrieved from http://www.sciencemag.org/content/289/5488/2309.abstract. https://doi.org/10.1126/science.289.5488.2309

Boeres, C., & Rebello, V. E. F. (2004). EasyGrid: Towards a framework for the automatic grid enabling of legacy MPI applications. *Concurrency and Computation: Practice and Experience*, *16*(5), 425–432. Retrieved from http://doi.wiley.com/10.1002/cpe.821. https://doi.org/10.1002/cpe.821

Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79. Retrieved from http://dl.acm.org/citation.cfm?doid=2723872.2723882http://arxiv.org/abs/1410.0846. https://doi.org/10.1145/2723872.2723882

Bonnet, P., Goëau, H., Hang, S. T., Lasseck, M., Šulc, M., Malécot, V., … Joly, A. (2018). Plant identification: Experts vs. machines in the era of deep learning. In *Multimedia tools and applications for environmental & biodiversity informatics* (pp. 131–149). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-76445-08. https://doi.org/10.1007/978-3-319-76445-08

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0304380013005917. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Borregaard, M. K., & Hart, E. M. (2016). Towards a more reproducible ecology. *Ecography*, *39*(4), 349–353. Retrieved from http://doi.wiley.com/10.1111/ecog.02493. https://doi.org/10.1111/ecog.02493

Borrett, S. R., Moody, J., & Edelmann, A. (2014). The rise of network ecology: Maps of the topic diversity and scientific collaboration. *Ecological Modelling*, *293*, 111–127. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0304380014001136. https://doi.org/10.1016/j.ecolmodel.2014.02.019

Borsch, T., Stevens, A.-D., Häffner, E., Güntsch, A., Berendsohn, W. G., Appelhans, M., … Zizka, G. (2020). A complete digitization of German herbaria is possible, sensible and should be started now. *Research Ideas and Outcomes*, *6*, e50675. https://riojournal.com/article/50675/. https://doi.org/10.3897/rio.6.e50675

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., … Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*(1), 16. Retrieved from http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-16. https://doi.org/10.1186/1471-2105-14-16

Brandao, S., Silva, W., Silva, L., Fagundes, V., de Mello, C., Zimbrao, G., & de Souza, J. (2009). *Analysis and Visualization of the Geographical Distribution of Atlantic Forest Bromeliads Species*. 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN: IEEE. pp. 375–380. Retrieved from http://ieeexplore.ieee.org/document/4938674/. https://doi.org/10.1109/CIDM.2009.4938674

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., … Turner, K. (2019). Computing environments for reproducibility: Capturing the "Whole Tale". *Future Generation Computer Systems*, *94*, 854–867. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0167739X17310695. https://doi.org/10.1016/j.future.2017.12.029

Bruce, T., Meirelles, P. M., Garcia, G., Paranhos, R., Rezende, C. E., de Moura, R. L., … Thompson, F. L. (2012). Abrolhos Bank reef health evaluated by means of water quality, microbial diversity, benthic cover, and fish biomass data. *PLoS ONE*, *7*(6), e36687. https://doi.org/10.1371/journal.pone.0036687

Burrell, A. S., Disotell, T. R., & Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution*, *79*, 35–44. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0047248414002619. https://doi.org/10.1016/j.jhevol.2014.10.015

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, *23*(1), 99–112. Retrieved from http://doi.wiley.com/10.1111/geb.12102. https://doi.org/10.1111/geb.12102

Campbell, K. L., & Hofreiter, M. (2012). New life for ancient DNA. *Scientific American*, *307*(2), 46–51. Retrieved from http://www.natu/doifinder/10.1038/scientificamerican0812-46. https://doi.org/10.1038/scientificamerican0812-46

Candela, L., Castelli, D., Coro, G., Pagano, P., & Sinibaldi, F. (2016). Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*, *28*(4), 1056–1079. Retrieved from http://doi.wiley.com/10.1002/cpe.3030. https://doi.org/10.1002/cpe.3030

Canhos, D. A. L., Sousa-Baena, M. S., de Souza, S., Maia, L. C., Stehmann, J. R., Canhos, V. P., … Peterson, A. T. (2015). The importance of biodiversity E-infrastructures for Megadiverse Countries. *PLOS Biology*, *13*(7), e1002204. Retrieved from http://dx.plos.org/10.1371/journal.pbio.1002204. https://doi.org/10.1371/journal.pbio.1002204

Cao, Y., Jones, C., Cuevas-Vicenttín, V., Jones, M. B., Ludäscher, B., McPhillips, T, … Wei, Y. (2016). *DataONE: A Data Federation with Provenance Support*. Provenance and Annotation of Data and Processes. IPAW 2016. Lecture Notes in Computer Science, Springer. Vol. 9672, McLean, pp. 230–234. Retrieved from http://link.springer.com/10.1007/978-3-319-40593-3_28 https://doi.org/10.1007/978-3-319-40593_28

Carata, L., Akoush, S., Balakrishnan, N., Bytheway, T., Sohan, R., Selter, M., & Hopper, A. (2014). A primer on provenance. *Communications of the ACM*, *57*(5), 52–60. Retrieved from http://dl.acm.org/gateway.cfm?id=2596628type=html. https://doi.org/10.1145/2596628

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., … Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), 59–67. Retrieved from http://www.nature.com/nature/journal/v486/n7401/full/nature11148.html?WT.ec_id=NATURE-20120607. https://doi.org/10.1038/nature11148

Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., & Joly, A. (2017). Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary Biology*, *17*(1), 181. Retrieved from http://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-017-1014-z. https://doi.org/10.1186/s12862-017-1014-z

Carranza-Rojas, J., Joly, A., Goëau, H., Mata-Montero, E., & Bonnet, P. (2018). Automated identification of herbarium specimens at different taxonomic levels. In *Multimedia tools and applications for environmental & biodiversity informatics* (pp. 151–167). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-76445-0_9. https://doi.org/10.1007/978-3-319-76445-09

Convention on Biological Diversity—CBD. (1992). *Text of the convention*. Retrieved from https://www.cbd.int/convention/text/default.shtml.

Chame, M., Barbosa, H. J. C., Gadelha, L. M. R., Augusto, D. A., Krempser, E., & Abdalla, L. (2019). SISS-geo: Leveraging citizen science to monitor wildlife health risks in Brazil. *Journal of Healthcare Informatics Research*, *3*(4), 414–440. Retrieved from http://link.springer.com/10.1007/s41666-019-00055-2. https://doi.org/10.1007/s41666-019-00055-2

Chapin, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., … Díaz, S. (2000). Consequences of changing biodiversity. *Nature*, *405*(6783), 234–242. Retrieved from https://doi.org/10.1038/35012241. https://doi.org/10.1038/35012241

Chapman, A. D. (2005). *Principles and methods of data cleaning—Primary species and species-occurence data* (Technical Report). Global Biodiversity Information Facility. Copenhagen, Denmark. https://www.gbif.org/document/80528/principles-and-methods-of-data-cleaning-primary-species-and-species-occurrence-data

Chard, K., Dart, E., Foster, I., Shifflett, D., Tuecke, S., & Williams, J. (2018). The modern research data portal: A design pattern for networked, data-intensive science. *PeerJ Computer Science*, *4*, e144. Retrieved from https://peerj.com/articles/cs-144. https://doi.org/10.7717/peerj-cs.144

Chase, J., & Leibold, M. (2003). *Ecological niches: Linking classical and contemporary approaches*, Chicago, IL: University of Chicago Press.

Chen, G., Han, T. X., He, Z., Kays, R., & Forrester, T. (2014). *Deep Convolutional Neural Network Based Species Recognition for Wild Animal Monitoring*. 2014 IEEE International Conference on Image Processing (ICIP), Paris, France: IEEE. pp. 858–862. Retrieved from http://ieeexplore.ieee.org/document/7025172/ https://doi.org/10.1109/ICIP.2014.7025172

Chen, Y. (2009). Conservation biogeography of the snake family Colubridae of China. *North-Western Journal of Zoology*, *5*(2), 251–262.

Chrisman, N. R. (1984). Part 2: Issues and problems relating to cartographic data use, exchange and transfer: The role of quality information in the long-term functioning of a geographic information system. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *21*(2–3), 79–88. Retrieved from https://utpjournals.press/doi/10.3138/7146–4332-6J78-0671. https://doi.org/10.3138/7146-4332-6J78-0671

Ciarleglio, M., Wesley Barnes, J., & Sarkar, S. (2009). ConsNet: New software for the selection of conservation area networks with spatial and multi-criteria analyses. *Ecography*, *32*(2), 205–209. Retrieved from http://doi.wiley.com/10.1111/j.1600-0587.2008.05721.x. https://doi.org/10.1111/j.1600-0587.2008.05721.x

Clark, M., Roberts, D., & Clark, D. (2005). Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, *96*(3–4), 375–398. Retrieved from http://www.sciencedirect.com/science/article/pii/S0034425705001082. https://doi.org/10.1016/j.rse.2005.03.009

Cobos, M. E., Peterson, A. T., Barve, N., & Osorio-Olvera, L. (2019). kuenm: An R package for detailed development of ecological niche models using Maxent. *PeerJ*, *7*, e6281. Retrieved from https://peerj.com/articles/6281. https://doi.org/10.7717/peerj.6281

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., … Blanchet, C. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, *75*, 284–298. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0167739X17300316. https://doi.org/10.1016/j.future.2017.01.012

Costa, J., Peterson, A. T., & Beard, C. B. (2002). Ecologic niche modeling and differentiation of populations of Triatoma brasiliensis neiva, 1911, the most important Chagas' disease vector in northeastern Brazil (hemiptera, reduviidae, triatominae). *The American Journal of Tropical Medicine and Hygiene*, *67*(5), 516–520. Retrieved from http://www.ajtmh.org/content/journals/10.4269/ajtmh.2002.67.516. https://doi.org/10.4269/ajtmh.2002.67.516

Couto-Lima, D., Madec, Y., Bersot, M. I., Campos, S. S., Motta, M. D. A., dos Santos, F. B., … Failloux, A.-B. (2017). Potential risk of re-emergence of urban transmission of Yellow Fever virus in Brazil facilitated by competent Aedes populations. *Scientific Reports*, *7*(1), 4848. Retrieved from http://www.nature.com/articles/s41598-017-05186-3. https://doi.org/10.1038/s41598-017-05186-3

Couvares, P., Kosar, T., Roy, A., Weber, J., & Wenger, K. (2007). Workflow management in Condor. In *Workflows for e-Science* (pp. 357–375). England: Springer. http://link.springer.com/10.1007/978-1-84628-757-2_22

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. Retrieved from http://doi.wiley.com/10.1890/07-0539.1. https://doi.org/10.1890/07-0539.1

da Silva, L. A. E., de Fraga, C. N., de Almeida, T. M. H., Gonzalez, M., Lima, R. O., da Rocha, M. S., … Forzza, R. C. (2017). Jabot - Sistema de Gerenciamento de Coleções Botânicas: A experiência de uma década de desenvolvimento e avanços. *Rodriguésia*, *68*(2), 391–410. Retrieved from http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2175-78602017000200391&lng=pt&tlng=pt. https://doi.org/10.1590/2175-7860201768208

Dalcin, E. C. (2005). *Data Quality Concepts and Techniques Applied to Taxonomic Databases* (Unpublished doctoral dissertation). University of Southampton. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.427422

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., … Davis, C. C. (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, *217*(2), 939–955. http://dx.doi.org/10.1111/nph.14855

de Andrade, A. F. A., Velazco, S. J. E., & De Marco Júnior, P. (2020). ENMTML: An R package for a straightforward construction of complex ecological niche models. *Environmental Modelling & Software*, *125*, 104615. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1364815219310424. https://doi.org/10.1016/j.envsoft.2019.104615

de Siqueira, M. F., & Peterson, A. T. (2003). Consequences of global climate change for geographic distribu- tions of cerrado tree species. *Biota Neotropica*, *3*(2), 1–14. https://doi.org/10.1590/S1676-06032003000200005

de Siracusa, P. C., Gadelha, L. M. R., & Ziviani, A. (2020). New perspectives on analysing data from biological collections based on social network analytics. *Scientific Reports*, *10*(1), 3358. Retrieved from http://www.nature.com/articles/s41598-020-60134-y. https://doi.org/10.1038/s41598-020-60134-y

Debeljak, M., Poljanec, A., & Ženko, B. (2014). Modelling forest growing stock from inventory data: A data mining approach. *Ecological Indicators*, *41*, 30–39. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1470160X14000181. https://doi.org/10.1016/j.ecolind.2014.01.010

Deelman, E., Gannon, D., Shields, M., & Taylor, I. (2009). Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, *25*(5), 528–540. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167739X08000861. https://doi.org/10.1016/j.future.2008.06.012

Diniz-Filho, J. A. F., Mauricio Bini, L., Fernando Rangel, T., Loyola, R. D., Hof, C., Nogués-Bravo, D., & Araújo, M. B. (2009). Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, *32*(6), 897–906. Retrieved from http://doi.wiley.com/10.1111/j.1600-0587.2009.06196.x. https://doi.org/10.1111/j.1600-0587.2009.06196.x

Dlamini, W. M. (2011). A data mining approach to predictive vegetation mapping using probabilistic graphical models. *Ecological Informatics*, *6*(2), 111–124. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1574954111000045. https://doi.org/10.1016/j.ecoinf.2010.12.005

Dong, X. L., & Rekatsinas, T. (2018). Data integration and machine learning. *Proceedings of the VLDB Endowment*, *11*(12), 2094–2097. Retrieved from http://dl.acm.org/citation.cfm?doid=3229863.3275606. https://doi.org/10.14778/3229863.3229876

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. Retrieved from http://doi.wiley.com/10.1111/j.1600-0587.2012.07348.x. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dormann, C. F., McPherson, J. M., B. Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., … Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. Retrieved from http://doi.wiley.com/10.1111/j.2007.0906-7590.05171.x. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Dou, L., Cao, G., Morris, P., Morris, R., Lud¨ascher, B., Macklin, J., & Hanken, J. (2012). Kurator: A kepler package for data curation workflows. *Procedia Computer Science*, *9*, 1614–1619. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877050912002980. https://doi.org/10.1016/j.procs.2012.04.177

Drew, L. W. (2011). Are we losing the science of taxonomy? *BioScience*, *61*(12), 942–946. Retrieved from https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2011.61.12.4. https://doi.org/10.1525/bio.2011.61.12.4

Droege, G., Barker, K., Astrin, J. J., Bartels, P., Butler, C., Cantrill, D., … Seberg, O. (2014). The global genome biodiversity network (GGBN) data portal. *Nucleic Acids Research*, *42* (Database issue, D607–D612. Retrieved from http://nar.oxfordjournals.org/content/early/2013/11/19/nar.gkt928. https://doi.org/10.1093/nar/gkt928

Droege, G., Barker, K., Seberg, O., Coddington, J., Benson, E., Berendsohn, W. G., … Zhou, X. (2016). The Global Genome Biodiversity Network (GGBN) data standard specification. *Database: The Journal of Biological Databases and Curation*, *2016*, baw125. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27694206http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5045859. https://doi.org/10.1093/database/baw125

Edwards, J. L. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, *289*(5488), 2312–2314. Retrieved from http://www.sciencemag.org/content/289/5488/2312.abstract. https://doi.org/10.1126/science.289.5488.2312

Edwards, J. L. (2004). Research and societal benefits of the global biodiversity information facility. *BioScience*, *54*(6), 486. Retrieved from http://bioscience.oxfordjournals.org/content/54/6/485.full. https://doi.org/10.1641/0006-3568(2004)054[0486:RASBOT]2.0.CO;2

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. Retrieved from http://doi.wiley.com/10.1111/j.2006.0906-7590.04596.x. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.110308.120159. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. Retrieved from http://doi.wiley.com/10.1111/j.1365-2656.2008.01390.x. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Eme, L., Spang, A., Lombard, J., Stairs, C. W., & Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, *15*(12), 711–723. Retrieved from http://www.nature.com/doifinder/10.1038/nrmicro.2017.133. https://doi.org/10.1038/nrmicro.2017.133

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, *41*(2), 263–274. Retrieved from http://doi.wiley.com/10.1111/j.0021-8901.2004.00881.x. https://doi.org/10.1111/j.0021-8901.2004.00881.x

Estrada-Peña, A., Ostfeld, R. S., Peterson, A. T., Poulin, R., & de la Fuente, J. (2014). Effects of environmental change on zoonotic disease risk: An ecological primer. *Trends in Parasitology*, *30*(4), 205–214. Retrieved from http://www.sciencedirect.com/science/article/pii/S1471492214000324. https://doi.org/10.1016/j.pt.2014.02.003

Faust, K. (2011). Animal social networks. In *The SAGE handbook of social network analysis* (pp. 148–166). England: SAGE Publications.

Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, *86*(3), 158–168. Retrieved from http://www.esajournals.org/doi/abs/10.1890/0012-9623(2005)865B158:MTVOED5D2.0.CO3B2. https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C., & Papeş, M. (2019). A checklist for maximizing reproducibility of ecological niche models. *Nature Ecology & Evolution*, *3*, 1382–1395. http://www.nature.com/articles/s41559-019-0972-5. https://doi.org/10.1038/s41559-019-0972-5

Fernández, N., Ferrier, S., Navarro, L. M., & Pereira, H. M. (2020). Essential biodiversity variables: Integrating in-situ observations and remote sensing through modeling. In *Remote sensing of plant biodiversity* (pp. 485–501). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-030-33157-318. https://doi.org/10.1007/978-3-030-33157-318

Filer, D. (2013). *BRAHMS—botanical research and herbarium management system: Training guide and introductory course*, Oxford, England: University of Oxford.

Flügge, A. J., Olhede, S. C., & Murrell, D. J. (2014). A method to detect subcommunities from multivariate spatial associations. *Methods in Ecology and Evolution*, *5*(11), 1214–1224. Retrieved from http://doi.wiley.com/10.1111/2041-210X.12295. https://doi.org/10.1111/2041-210X.12295

Forzza, R. C., Baumgratz, J. F. A., Bicudo, C. E. M., Canhos, D. A. L., Carvalho, A. A., Coelho, M. A. N., … Zappi, D. C. (2012). New Brazilian floristic list highlights conservation challenges. *BioScience*, *62*(1), 39–45. Retrieved from http://bioscience.oxfordjournals.org/content/62/1/39.full. https://doi.org/10.1525/bio.2012.62.1.8

Francini-Filho, R. B., Asp, N. E., Siegle, E., Hocevar, J., Lowyck, K., D'Avila, N., … Thompson, F. L. (2018). Perspectives on the great Amazon Reef: Extension, biodiversity, and threats. *Frontiers in Marine Science*, *5*, 142. http://journal.frontiersin.org/article/10.3389/fmars.2018.00142/full. https://doi.org/10.3389/fmars.2018.00142

Freire, J., & Chirigati, F. (2018). Provenance and the different flavors of computational reproducibility. *Bulletin of the Technical Committee on Data Engineering*, *41*(1), 15–26 Retrieved from http://sites.computer.org/debull/A18mar/p15.pdf

Freire, J., Koop, D., Santos, E., & Silva, C. (2008). Provenance for computational tasks: A survey. *Computing in Science & Engineering*, *10*(3), 11–21. Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4488060. https://doi.org/10.1109/MCSE.2008.79

Freire, J., Silva, C. T., Callahan, S. P., Santos, E., Scheidegger, C. E., & Vo, H. T. (2006). Managing rapidly-evolving scientific workflows. In *Provenance and Annotation of Data. IPAW 2006*, Lecture Notes in Computer Science, 4145, 10–18. Berlin, Heidelberg: Springer.

Fretwell, P. T., LaRue, M. A., Morin, P., Kooyman, G. L., Wienecke, B., Ratcliffe, N., … Trathan, P. N. (2012). An Emperor Penguin population estimate: The first global, synoptic survey of a species from space. *PLoS ONE*, *7*(4), e33751. Retrieved from http://dx.plos.org/10.1371/journal.pone.0033751. https://doi.org/10.1371/journal.pone.0033751

Fujioka, E., Kot, C. Y., Wallace, B. P., Best, B. D., Moxley, J., Cleary, J., … Halpin, P. N. (2014). Data integration for conservation: Leveraging multiple data types to advance ecological assessments and habitat modeling for marine megavertebrates using OBIS-SEAMAP. *Ecological Informatics*, *20*, 13–26. Retrieved from http://www.sciencedirect.com/science/article/pii/S1574954114000041. https://doi.org/10.1016/j.ecoinf.2014.01.003

Gadelha, L., Guimarães, P., Moura, A. M., Drucker, D. P., Dalcin, E., Gall, G, … Leo, W. V. (2014). *SiBBr: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira*. In G. D. Garcia, G. B. Gregoracci, E. D. O. Santos, P. M. Meirelles, G. G. Z. Silva, R. Edwards. Viii Brazilian e-Science Workshop (BRESCI 2014). Proceedings of xxxiv Congress of the Brazilian Computer Society. Brasília, Brazil. https://sol.sbc.org.br/index.php/bresci/article/view/10477

Garcia, G. D., Gregoracci, G. B., Santos, E. de O., Meirelles, P. M., Silva, G. G. Z., Edwards, R., … Thompson, F. L. (2013). Metagenomic analysis of healthy and white plague-affected Mussismilia braziliensis corals. *Microbial Ecology*, *65*(4), 1076–1086. https://doi.org/10.1007/s00248-012-0161-4

Gerlach, R., Blaa, D., Chamanara, J., Hohmuth, M., Navabpour, N., Thiel, S., & König-Ries, B. (2015). *BEXIS 2: A Platform for Managing Heterogeneous Biodiversity Data and Projects*. Tdwg 2015 Annual Conference. Nairobi, Kenya

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, *13*(5), 533–549. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/0305054886900481. https://doi.org/10.1016/0305-0548(86)90048-1

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., … Stanzione, D. (2011). The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Frontiers in Plant Science*, *2*, 34. http://dx.doi.org/10.3389/fpls.2011.00034.

Golding, N., August, T. A., Lucas, T. C. D., Gavaghan, D. J., van Loon, E. E., & McInerny, G. (2018). The ZOON R package for reproducible and shareable species distribution modelling. *Methods in Ecology and Evolution*, *9*(2), 260–268. Retrieved from http://doi.wiley.com/10.1111/2041-210X.12858. https://doi.org/10.1111/2041-210X.12858

Gomes, V. H. F., IJff, S. D., Raes, N., Amaral, I. L., Salomão, R. P., de Souza Coelho, L., … ter Steege, H. (2018). Species distribution modelling: Contrasting presence-only models with plot abundance data. *Scientific Reports*, *8*(1), 1003. Retrieved from http://www.nature.com/articles/s41598-017-18927-1. https://doi.org/10.1038/s41598-017-18927-1

Gomez Villa, A., Salazar, A., & Vargas, F. (2017). Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, *41*, 24–32. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1574954116302047. https://doi.org/10.1016/j.ecoinf.2017.07.004

Grassle, J. (2000). The ocean biogeographic information system (OBIS): An on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. *Oceanography*, *13*(3), 5–7.

Groom, Q. J., O'Reilly, C., & Humphrey, T. (2014). Herbarium specimens reveal the exchange network of British and Irish botanists, 1856–1932. *New Journal of Botany*, *4*(2), 95–103. Retrieved from http://www.tandfonline.com/doi/full/10.1179/2042349714Y.0000000041. https://doi.org/10.1179/2042349714Y.0000000041

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., & Peterson, A. T. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, *77*(4), 615–630. Retrieved from http://doi.wiley.com/10.1890/06-1060.1. https://doi.org/10.1890/06-1060.1

Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., R¨opert, D., Casino, A., … Triebel, D. (2017). Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database*, *2017*, bax003https://academic.oup.com/database/article/doi/10.1093/database/bax003/3053443. https://doi.org/10.1093/database/bax003

Guralnick, R., & Hill, A. (2009). Biodiversity informatics: Automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics (Oxford, England)*, *25*(4), 421–428. Retrieved from http://bioinformatics.oxfordjournals.org/content/25/4/421. https://doi.org/10.1093/bioinformatics/btn659

Guralnick, R. P., Wieczorek, J., Beaman, R., & Hijmans, R. J. (2006). BioGeomancer: Automated georeferencing to map the world's biodiversity data. *PLoS Biology*, *4*(11), e381. Retrieved from http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040381. https://doi.org/10.1371/journal.pbio.0040381

Guru, S., Hanigan, I. C., Nguyen, H. A., Burns, E., Stein, J., Blanchard, W., … Clancy, T. (2016). Development of a cloud-based platform for reproducible science: A case study of an IUCN red list of ecosystems assessment. *Ecological Informatics*, *36*, 221–230. Retrieved from https://www.sciencedirect.com/science/article/pii/S1574954116301182. https://doi.org/10.1016/J.ECOINF.2016.08.003

Gwinn, N. E., & Rinaldo, C. (2009). The biodiversity heritage library: Sharing biodiversity literature with the world. *IFLA Journal*, *35*(1), 25–34. Retrieved from http://ifl.sagepub.com/content/35/1/25.short. https://doi.org/10.1177/0340035208102032

Hale, J. S., Li, L., Richardson, C. N., & Wells, G. N. (2017). Containers for Portable, productive, and performant scientific computing. *Computing in Science & Engineering*, *19*(6), 40–50. Retrieved from http://ieeexplore.ieee.org/document/7933304/. https://doi.org/10.1109/MCSE.2017.2421459

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed. Waltham, Massachusetts). Morgan Kaufmann.

Hardisty, A., Roberts, D., Addink, W., Aelterman, B., Agosti, D., Amaral-Zettler, L., … Young, F. (2013). A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, *13*(1), 16. Retrieved from http://www.biomedcentral.com/1472-6785/13/16. https://doi.org/10.1186/1472-6785-13-16

Hardisty, A. R., Bacall, F., Beard, N., Balc´azar-Vargas, M.-P., Balech, B., Barcza, Z., … Yilmaz, P. (2016). BioVeL: A virtual laboratory for data analysis and modelling in biodiversity science and ecology. *BMC Ecology*, *16*(1), 49. Retrieved from http://bmcecol.biomedcentral.com/articles/10.1186/s12898-016-0103-y. https://doi.org/10.1186/s12898-016-0103-y

Hardisty, A. R., Michener, W. K., Agosti, D., Alonso García, E., Bastin, L., Belbin, L., … Kissling, W. D. (2019). The Bari Manifesto: An interoperability framework for essential biodiversity variables. *Ecological Informatics*, *49*, 22–31. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1574954118301961. https://doi.org/10.1016/j.ecoinf.2018.11.003

Harjes, J., Link, A., Weibulat, T., Triebel, D., & Rambold, G. (2020). FAIR digital objects in environmental and life sciences should comprise workflow operation design data and method information for repeatability of study setups and reproducibility of results. *Database*, *2020*, baaa059. https://academic.oup.com/database/article/doi/10.1093/database/baaa059/5894776. https://doi.org/10.1093/database/baaa059

Haston, E., Cubey, R., Pullan, M., Atkins, H., & Harris, D. (2012). Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. *ZooKeys*, *209*, 93–102. Retrieved from http://zookeys.pensoft.net/articles.php?id=2920. https://doi.org/10.3897/zookeys.209.3121

He, Y., Preece, J., Hammock, J., Butler, B., & Pauw, D. (2015). *Understanding Data Providers in a Global Scientific Data Hub*. Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative work & Social Computing—CSCW'15 Companion, ACM Press, New York, NY. Vancouver, Canada. pp. 215–218. Retrieved from http://dl.acm.org/citation.cfm?id=2685553.2699010 https://doi.org/10.1145/2685553.2699010

He, Z., Kays, R., Zhang, Z., Ning, G., Huang, C., Han, T. X., … McShea, W. (2016). Visual informatics tools for supporting large-scale collaborative wildlife monitoring with citizen scientists. *IEEE Circuits and Systems Magazine*, *16*(1), 73–86. Retrieved from http://ieeexplore.ieee.org/document/7404334/. https://doi.org/10.1109/MCAS.2015.2510200

Heberling, J. M., & Isaac, B. L. (2018). iNaturalist as a tool to expand the research value of museum specimens. *Applications in Plant Sciences*, *6*(11), e01193. http://dx.doi.org/10.1002/aps3.1193

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. Retrieved from http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2002.2218. https://doi.org/10.1098/rspb.2002.2218

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Heikkinen, R. K., Luoto, M., Virkkala, R., Pearson, R. G., & Körber, J.-H. (2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, *16*(6), 754–763. Retrieved from http://doi.wiley.com/10.1111/j.1466-8238.2007.00345.x. https://doi.org/10.1111/j.1466-8238.2007.00345.x

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

Hey, T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science (New York, N.Y.)*, *308*(5723), 817–821. Retrieved from http://science.sciencemag.org/content/308/5723/817.abstract. https://doi.org/10.1126/science.1110410

Higgins, D., Berkley, C., & Jones, M. (2002). *Managing Heterogeneous Ecological Data Using Morpho*. Proceedings 14th International Conference on Scientific and Statistical Database Management, Edinburgh, Scotland: IEEE Computer Society. pp. 69–76. Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1029707 https://doi.org/10.1109/SSDM.2002.1029707

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. Retrieved from http://doi.wiley.com/10.1002/joc.1276. https://doi.org/10.1002/joc.1276

Hobern, D., Apostolico, A., Arnaud, E., Bello, J. C., Canhos, D., Dubois, G., … Willoughby, S. (2013). *Global biodiversity information outlook—Delivering biodiversity knowledge in the information age* (Technical Report). GBIF Secretariat. Retrieved from http://www.biodiversityinformatics.org/download-gbio-report/

Hobern, D., Baptiste, B., Copas, K., Guralnick, R., Hahn, A., van Huis, E., … Wieczorek, J. (2019). Connecting data and expertise: A new alliance for biodiversity knowledge. *Biodiversity Data Journal*, *7*, e33679https://bdj.pensoft.net/article/33679/. https://doi.org/10.3897/BDJ.7.e33679

Hochachka, W. M., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., & Kelling, S. (2007). Data-mining discovery of pattern and process in ecological systems. *Journal of Wildlife Management*, *71*(7), 2427. Retrieved from http://www.bioone.org/perlserv/?request=get-abstract&doi=10.2193_2F2006-503. https://doi.org/10.2193/2006-503

Holden, H., & Ledrew, E. (1999). Hyperspectral identification of coral reef features. *International Journal of Remote Sensing*, *20*(13), 2545–2563. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/014311699211921. https://doi.org/10.1080/014311699211921

Holmes, D., & McCabe, M. (2002). *Improving Precision and Recall for Soundex Retrieval*. Proceedings. International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada: IEEE Computer Society. pp. 22–26. Retrieved from http://ieeexplore.ieee.org/document/1000354/ https://doi.org/10.1109/ITCC.2002.1000354

Holmes, E. C. (2008). Evolutionary history and phylogeography of human viruses. *Annual Review of Microbiology*, *62*(1), 307–328. Retrieved from http://www.annualreviews.org/doi/10.1146/annurev.micro.62.081307.162912. https://doi.org/10.1146/annurev.micro.62.081307.162912

Hooper, D. U., Adair, E. C., Cardinale, B. J., Byrnes, J. E. K., Hungate, B. A., Matulich, K. L., … O'Connor, M. I. (2012). A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature*, *486*(7401), 105–108. https://doi.org/10.1038/nature11118

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., … Yon Rhee, S. (2008). Big data: The future of biocuration. *Nature*, *455* (7209), 47–50. Retrieved from http://www.nature.com/articles/455047a. https://doi.org/10.1038/455047a

Ings, T. C., Montoya, J. M., Bascompte, J., Blüthgen, N., Brown, L., Dormann, C. F., … Woodward, G. (2009). Review: Ecological networks—Beyond food webs. *Journal of Animal Ecology*, *78*(1), 253–269. Retrieved from http://doi.wiley.com/10.1111/j.1365-2656.2008.01460.x. https://doi.org/10.1111/j.1365-2656.2008.01460.x

Inman-Narahari, F., Giardina, C., Ostertag, R., Cordell, S., & Sack, L. (2010). Digital data collection in forest dynamics plots. *Methods in Ecology and Evolution*, *1*(3), 274–279. Retrieved from http://doi.wiley.com/10.1111/j.2041-210X.2010.00034.x. https://doi.org/10.1111/j.2041-210X.2010.00034.x

Ivie, P., & Thain, D. (2018). Reproducibility in scientific computing. *ACM Computing Surveys*, *51*(3), 1–36. Retrieved from http://dl.acm.org/citation.cfm?doid=3212709.3186266. https://doi.org/10.1145/3186266

Jacoby, D. M., & Freeman, R. (2016). Emerging network-based tools in movement ecology. *Trends in Ecology & Evolution*, *31*(4), 301–314. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169534716000264. https://doi.org/10.1016/j.tree.2016.01.011

Jones, M. B., Schildhauer, M. P., Reichman, O., & Bowers, S. (2006). The New bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 519–544. Retrieved from http://www.annualreviews.org/eprint/DjakMtfpntHPCZbuBK5y/full/10.1146/annurev.ecolsys.37.091305.110031?utm_source=&utm_medium=environ&utm_campaign=eprint. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Kass, J. M., Vilela, B., Aiello-Lammens, M. E., Muscarella, R., Merow, C., & Anderson, R. P. (2018). Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolution*, *9*(4), 1151–1156. Retrieved from http://doi.wiley.com/10.1111/2041-210X.12945 10.1111/2041-210X.12945.

Keith, D. A., Rodríguez, J. P., Rodríguez-Clark, K. M., Nicholson, E., Aapala, K., Alonso, A., … Zambrano-Martínez, S. (2013). Scientific foundations for an IUCN red list of ecosystems. *PLoS ONE*, *8*(5), e62111. Retrieved from http://dx.plos.org/10.1371/journal.pone.0062111. https://doi.org/10.1371/journal.pone.0062111

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J., & Kreft, H. (2019). Biodiversity data integration—The significance of data resolution and domain. *PLOS Biology*, *17*(3), e3000183. Retrieved from http://dx.plos.org/10.1371/journal.pbio.3000183. https://doi.org/10.1371/journal.pbio.3000183

Koning, D., Sarkar, I. N., & Moritz, T. (2005). TaxonGrab: Extracting taxonomic names from text. *Biodiversity Informatics*, *2*, 79–82. https://journals.ku.edu/jbi/article/view/17. https://doi.org/10.17161/bi.v2i0.17

Köpcke, H., Thor, A., & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, *3*(1–2), 484–493. Retrieved from http://dl.acm.org/citation.cfm?doid=1920841.1920904. https://doi.org/10.14778/1920841.1920904

Kumar, J., Mills, R. T., Hoffman, F. M., & Hargrove, W. W. (2011). Parallel k-means clustering for quantitative Ecoregion delineation using large data sets. *Procedia Computer Science*, *4*, 1602–1611. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1877050911002316. https://doi.org/10.1016/j.procs.2011.04.173

La Salle, J., Williams, K. J., & Moritz, C. (2016). Biodiversity analysis in the digital era. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, *371*(1702), 534–547. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27481789http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4971189. https://doi.org/10.1098/rstb.2015.0337

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Wintle, B. A. (2014). Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, *23*(4), 504–515. Retrieved from http://doi.wiley.com/10.1111/geb.12138. https://doi.org/10.1111/geb.12138

Lanna, J., da Silva, L. A., Morim, M., Leitman, P., Queiroz, N., Filardi, F., … Forzza, R. (2018). Herbarium collection of the Rio de Janeiro Botanical Garden (RB), Brazil. *Biodiversity Data Journal*, *6*, e22757. Retrieved from https://bdj.pensoft.net/articles.php?id=22757. https://doi.org/10.3897/BDJ.6.e22757

Leader, B., Baca, Q. J., & Golan, D. E. (2008). Protein therapeutics: A summary and pharmacological classification. *Nature Reviews Drug Discovery*, *7*(1), 21–39. Retrieved from http://www.nature.com/articles/nrd2399. https://doi.org/10.1038/nrd2399

Lee, C. T., & Amaro, R. E. (2018). Exascale computing: A new dawn for computational biology. *Computing in Science & Engineering*, *20*(5), 18–25. Retrieved from https://ieeexplore.ieee.org/document/8452060/. https://doi.org/10.1109/MCSE.2018.05329812

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, *10*(8), 707–710.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0957417412003077. https://doi.org/10.1016/j.eswa.2012.02.063

Liew, C. S., Atkinson, M. P., Galea, M., Ang, T. F., Martin, P., & Hemert, J. I. V. (2016). Scientific workflows: Moving across paradigms. *ACM Computing Surveys*, *49*(4), 1–39. Retrieved from http://dl.acm.org/citation.cfm?doid=3022634.3012429. https://doi.org/10.1145/3012429

Lindeman, R. L. (1942). The trophic-dynamic aspect of ecology. *Ecology*, *23*(4), 399–417. Retrieved from http://doi.wiley.com/10.2307/1930126. https://doi.org/10.2307/1930126

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*(3), 385–393. Retrieved from http://doi.wiley.com/10.1111/j.0906-7590.2005.03957.x. https://doi.org/10.1111/j.0906-7590.2005.03957.x

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, *34*(2), 232–243. Retrieved from http://doi.wiley.com/10.1111/j.1600-0587.2010.06354.x. https://doi.org/10.1111/j.1600-0587.2010.06354.x

Liu, J., Pacitti, E., Valduriez, P., & Mattoso, M. (2015). A survey of data-intensive scientific workflow management. *Journal of Grid Computing*, *13*(4), 457–493. Retrieved from http://link.springer.com/10.1007/s10723-015-9329-8. https://doi.org/10.1007/s10723-015-9329-8

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. Retrieved from http://doi.wiley.com/10.1111/j.1466-8238.2007.00358.x. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lomolino, M. (2004). Conservation biogeography. In *Frontiers in biogeography: New directions in the geography of nature* (pp. 293–296). Sunderland, MA. Sinauer.

Lorena, A. C., Jacintho, L. F., Siqueira, M. F., Giovanni, R. D., Lohmann, L. G., de Carvalho, A. C., & Yamamoto, M. (2011). Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, *38*(5), 5268–5275. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0957417410011759. https://doi.org/10.1016/j.eswa.2010.10.031

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., … Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, *18*(10), 1039–1065. Retrieved from http://doi.wiley.com/10.1002/cpe.994. https://doi.org/10.1002/cpe.994

Magnusson, W., Braga-Neto, R., Pezzini, F., Baccaro, F., Bergallo, H., Penha, J., … Pontes, A. R. M. (2013). *Biodiversity and integrated environmental monitoring*. Attema Editorial. Retrieved from http://ppbio.inpa.gov.br/sites/default/files/Biodiversidadeemonitoramentoambientalintegrado.pdf

Martijn, J., Vosseberg, J., Guy, L., Offre, P., & Ettema, T. J. G. (2018). Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature*, *557*(7703), 101–105. Retrieved from http://www.nature.com/articles/s41586-018-0059-5. https://doi.org/10.1038/s41586-018-0059-5

Mattoso, M., Werner, C., Travassos, G. H., Braganholo, V., Ogasawara, E., Oliveira, D., … Murta, L. (2010). Towards supporting the life cycle of large scale scientific experiments. *International Journal of Business Process Integration and Management*, *5*(1), 79–92. http://www.inderscienceonline.com/doi/abs/10.1504/IJBPIM.2010.033176

McNeill, J. (2012). *International code of nomenclature for algae, fungi and plants (Melbourne code)*. Adopted by the Eighteenth International Botanical Congress Melbourne. Koeltz Scientific Books.

Meirelles, P. M., Amado-Filho, G. M., Pereira-Filho, G. H., Pinheiro, H. T., de Moura, R. L., Joyeux, J.-C., … Thompson, F. L. (2015). Baseline assessment of mesophotic reefs of the Vitória-Trindade Seamount Chain based on water quality, microbial diversity, benthic cover and fish biomass data. *PLOS ONE*, *10*(6), e0130084. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130084. https://doi.org/10.1371/journal.pone.0130084

Meng, H., Kommineni, R., Pham, Q., Gardner, R., Malik, T., & Thain, D. (2015). An invariant framework for conducting reproducible computational science. *Journal of Computational Science*, *9*, 137–142. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877750315000502. https://doi.org/10.1016/j.jocs.2015.04.012

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., … Edwards, R. A. (2008). The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, *9*, 386. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2563014&tool=pmcentrez&rendertype=abstract. https://doi.org/10.1186/1471-2105-9-386

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., … Vieglais, D. A. (2012). Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, *11*, 5–15. Retrieved from http://www.sciencedirect.com/science/article/pii/S1574954111000768. https://doi.org/10.1016/j.ecoinf.2011.08.007

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85–93. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169534711003399. https://doi.org/10.1016/j.tree.2011.11.016

Michener, W. K., Porter, J., Servilla, M., & Vanderbilt, K. (2011). Long term ecological research and information management. *Ecological Informatics*, *6*(1), 13–24. Retrieved from http://www.sciencedirect.com/science/article/pii/S1574954110001159. https://doi.org/10.1016/j.ecoinf.2010.11.005

Miles, S., Wong, S. C., Fang, W., Groth, P., Zauner, K.-P., & Moreau, L. (2007). Provenance-based validation of e-science experiments. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(1), 28–38. Retrieved from http://www.sciencedirect.com/science/article/pii/S1570826806000564. https://doi.org/10.1016/j.websem.2006.11.003

Miller, R. J. (2018). Open data integration. *Proceedings of the VLDB Endowment*, *11*(12), 2130–2139. Retrieved from http://www.vldb.org/pvldb/vol11/p2130-miller.pdf. https://doi.org/10.14778/3229863.3240491

Mills, R. T., Hoffman, F. M., Kumar, J., & Hargrove, W. W. (2011). Cluster analysis-based approaches for geospatiotemporal data mining of massive data sets for identification of forest threats. *Procedia Computer Science*, *4*, 1612–1621. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1877050911002328. https://doi.org/10.1016/j.procs.2011.04.174

Mondelli, M. L., Townsend Peterson, A., & Gadelha, L. M. R. (2019). *Exploring Reproducibility and FAIR Principles in Data Science Using Ecological Niche Modeling as a Case Study*. Advances in Conceptual Modeling. ER 2019. Lecture Notes in Computer Science, Salvador, Brazil:

Springer. Vol. 11787, pp. 23–33. Retrieved from http://link.springer.com/10.1007/978-3-030-34146-6_3 https://doi.org/10.1007/978-3-030-34146-63

Moreau, L., Groth, P., Cheney, J., Lebo, T., & Miles, S. (2015). The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web*, *35*, 235–257. Retrieved from http://www.sciencedirect.com/science/article/pii/S1570826815000177. https://doi.org/10.1016/j.websem.2015.04.001

Moreira-Soto, A., Torres, M. C., Lima de Mendonça, M. C., Mares-Guia, M. A., Damasceno dos Santos Rodrigues, C., Fabri, A., … Bispo de Filippis, A. M. (2018). Evidence for multiple sylvatic transmission cycles during the 2016-2017 yellow fever virus outbreak, Brazil. *Clinical Microbiology and Infection.* Retrieved from https://www.sciencedirect.com/science/article/pii/S1198743X18301447, *24*, 1019.e1–1019.e4. https://doi.org/10.1016/J.CMI.2018.01.026

Morisette, J. T., Jarnevich, C. S., Holcombe, T. R., Talbert, C. B., Ignizio, D., Talbert, M. K., … Young, N. E. (2013). VisTrails SAHM: Visualization and workflow management for species habitat modeling. *Ecography*, *36*(2), 129–135. Retrieved from http://doi.wiley.com/10.1111/j.1600-0587.2012.07815.x. https://doi.org/10.1111/j.1600-0587.2012.07815.x

Morris, P. J., Hanken, J., Lowery, D., Ludäscher, B., Macklin, J., McPhillips, T., … Zhang, Q. (2018). Kurator: Tools for improving fitness for use of biodiversity data. *Biodiversity Information Science and Standards*, *2*, e26539. Retrieved from https://biss.pensoft.net/articles.php?id=26539. https://doi.org/10.3897/biss.2.26539

Morris, R. A., Barve, V., Carausu, M., Chavan, V., Cuadra, J., Freeland, C., … Whitbread, G. (2013). Discovery and publishing of primary biodiversity data associated with multimedia resources: The Audubon core strategies and approaches. *Biodiversity Informatics*, *8*(2), 185–197. https://ojsprdap.vm.ku.edu/index.php/jbi/article/view/4117. https://doi.org/10.17161/bi.v8i2.4117

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., … Raghavendra, V. (2018). *Deep Learning for Entity Matching*. Proceedings of the 2018 International Conference on Management of Datasigmod '18, ACM Press, New York, NY. pp. 19–34. Retrieved from http://dl.acm.org/citation.cfm?doid=3183713.3196926 https://doi.org/10.1145/3183713.3196926

Naimi, B., Skidmore, A. K., Groen, T. A., & Hamm, N. A. S. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, *38*(8), 1497–1509. Retrieved from http://doi.wiley.com/10.1111/j.1365-2699.2011.02523.x. https://doi.org/10.1111/j.1365-2699.2011.02523.x

Nargesian, F., Zhu, E., Pu, K. Q., & Miller, R. J. (2018). Table union search on open data. *Proceedings of the VLDB Endowment*, *11*(7), 813–825. Retrieved from http://dl.acm.org/citation.cfm?doid=3192965.3228339. https://doi.org/10.14778/3192965.3192973

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., … Purvis, A. (2015). Global effects of land use on local terrestrial biodiversity. *Nature*, *520*(7545), 45–50. Retrieved from https://doi.org/10.1038/nature14324. https://doi.org/10.1038/nature14324

Newman, M. (2010). *Networks: An introduction*, Oxford, England: Oxford University Press.

Nguyen, H. A., Bland, L., Roberts, T., Guru, S., Dinh, M., & Abramson, D. (2017). *A Computational Pipeline for the IUCN Risk Assessment for Meso-American Reef Ecosystem*. 2017 IEEE 13th International Conference on e-Science (e-Science). Auckland, New Zealand. pp. 286–294. Retrieved from http://ieeexplore.ieee.org/document/8109147/ https://doi.org/10.1109/eScience.2017.42

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, *115*(25), E5716–E5725. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1719367115. https://doi.org/10.1073/pnas.1719367115

O'Leary, M. A., & Kaufman, S. (2011). MorphoBank: Phylophenomics in the "cloud". *Cladistics*, *27*(5), 529–537. Retrieved from http://doi.wiley.com/10.1111/j.1096-0031.2011.00355.x. https://doi.org/10.1111/j.1096-0031.2011.00355.x

O'Tuama, E., Deck, J., Dröge, G., Döring, M., Field, D., Kottmann, R., … Yilmaz, P. (2012). Meeting Report: Hackathon-workshop on Darwin Core and MIxS standards alignment. *Standards in Genomic Sciences*, *7*(1), 166–170. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3570805&tool=pmcentrez&rendertype=abstract. https://doi.org/10.4056/sigs.3166513

Odum, H. T. (1956). Primary production in flowing waters. *Limnology and Oceanography*, *1*(2), 102–117. https://doi.org/10.4319/lo.1956.1.2.0102

Otegui, J., & Guralnick, R. P. (2016). The geospatial data quality REST API for primary biodiversity data. *Bioinformatics (Oxford, England)*, *32*(11), 1755–1757. https://doi.org/10.1093/bioinformatics/btw057

Owens, H. L., Campbell, L. P., Dornak, L. L., Saupe, E. E., Barve, N., Soberón, J., … Peterson, A. T. (2013). Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling*, *263*, 10–18. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0304380013002159. https://doi.org/10.1016/j.ecolmodel.2013.04.011

Page, R. D. M. (2008). Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, *9*(5), 345–354. Retrieved from http://bib.oxfordjournals.org/content/9/5/345.short. https://doi.org/10.1093/bib/bbn022

Paul, D., Mast, A. R., Riccardi, G., & Nelson, G. (2013). *iDigBio as a Resource for the Digitization of a Billion Biodiversity Research Specimens*. Tdwg 2013 Annual Conference. Florence, Italy. Retrieved from https://mbgocs.mobot.org/index.php/tdwg/2013/paper/view/377/0

Pearson, R. G. (2010). Species' sistribution modeling for conservation educators and practiotioners. *Lessons in Conservation*, *3*, 54–89.

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Townsend Peterson, A. (2006). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography*, *34*(1), 102–117. Retrieved from http://doi.wiley.com/10.1111/j.1365-2699.2006.01594.x. https://doi.org/10.1111/j.1365-2699.2006.01594.x

Pearson, R. G., Thuiller, W., Araújo, M. B., Martinez-Meyer, E., Brotons, L., McClean, C., … Lees, D. C. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography*, *33*(10), 1704–1711. https://doi.org/10.1111/j.1365-2699.2006.01460.x

Peña, C., & Malm, T. (2012). VoSeq: A voucher and DNA sequence web application. *PLoS ONE*, *7*(6), 1–4. https://doi.org/10.1371/journal.pone.0039071

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Pennington, D. D., Higgins, D., Peterson, A. T., Jones, M. B., Ludäscher, B., & Bowers, S. (2007). Ecological niche modeling using the Kepler workflow system. In I. J. Taylor, E. Deelman, D. B. Gannon, & M. Shields (Eds.), *Workflows for e-science* (pp. 91–108). London: Springer.

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., … Wegmann, M. (2013). Ecology. Essential biodiversity variables. *Science (New York, N.Y.)*, *339*(6117), 277–278. Retrieved from http://www.sciencemag.org/content/339/6117/277.full. https://doi.org/10.1126/science.1229931

Pergl, R., Hooft, R., Such´anek, M., Knaisl, V., & Slifka, J. (2019). "Data Stewardship Wizard": A tool bringing together researchers, data stewards, and data experts around data management planning. *Data Science Journal*, *18*(1), 59http://datascience.codata.org/articles/10.5334/dsj-2019-059/. https://doi.org/10.5334/dsj-2019-059

Peterson, A. T. (2006). Uses and requirements of ecological Niche models and related distributional models. *Biodiversity Informatics*, *3*, 59–72. https://journals.ku.edu/index.php/jbi/article/view/29. https://doi.org/10.17161/bi.v3i0.29

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*(1), 63–72. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0304380007006163. https://doi.org/10.1016/j.ecolmodel.2007.11.008

Peterson, A. T., & Robins, C. R. (2003). Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. *Conservation Biology*, *17*(4), 1161–1165. Retrieved from http://doi.wiley.com/10.1046/j.1523-1739.2003.02206.x. https://doi.org/10.1046/j.1523-1739.2003.02206.x

Peterson, A. T., & Soberón, J. (2017). Essential biodiversity variables are not global. *Biodiversity and Conservation*, *27*, 1277–1288. http://link.springer.com/10.1007/s10531-017-1479-5. https://doi.org/10.1007/s10531-017-1479-5

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions*, Princeton, NJ: Princeton University Press.

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T. P., Fernandez, M., … Geller, G. N. (2016). Framing the concept of satellite remote sensing essential biodiversity variables: Challenges and future directions. *Remote Sensing in Ecology and Conservation*, *2*(3), 122–131. http://doi.wiley.com/10.1002/rse2.15. https://doi.org/10.1002/rse2.15

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. Retrieved from http://www.sciencedirect.com/science/article/pii/S030438000500267X. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Piatetsky-Shapiro, G., & Frawley, W. (1989). *Knowledge Discovery in Databases*. IJCAI-89 Workshop Proceedings. Detroit, Michigan

Pino-Mejías, R., Cubiles-de-la Vega, M. D., Anaya-Romero, M., Pascual-Acosta, A., Jordán-López, A., & Bellinfante-Crocci, N. (2010). Predicting the potential habitat of oaks with data mining models and the R system. *Environmental Modelling & Software*, *25*(7), 826–836. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364815210000150. https://doi.org/10.1016/j.envsoft.2010.01.004

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. Retrieved from https://peerj.com/articles/175. https://doi.org/10.7717/peerj.175

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., … van Swaay, C. A. (2017). Global biodiversity monitoring: From data sources to essential biodiversity variables. *Biological Conservation*, *213*, 256–263. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0006320716302786. https://doi.org/10.1016/j.biocon.2016.07.014

Qiao, H., Peterson, A. T., Campbell, L. P., Soberón, J., Ji, L., & Escobar, L. E. (2016). NicheA: Creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, *39*(8), 805–813. Retrieved from http://doi.wiley.com/10.1111/ecog.01961. https://doi.org/10.1111/ecog.01961

Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, *6*(10), 1126–1136. Retrieved from http://doi.wiley.com/10.1111/2041-210X.12397. https://doi.org/10.1111/2041-210X.12397

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1890991&tool=pmcentrez&rendertype=abstract. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rees, T. (2014). Taxamatch, an algorithm for near ('Fuzzy') matching of scientific names in taxonomic databases. *PloS one*, *9*(9), e107510. Retrieved from http://dx.plos.org/10.1371/journal.pone.0107510. https://doi.org/10.1371/journal.pone.0107510

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science (New York, N.Y.)*, *331*(6018), 703–705. Retrieved from http://www.sciencemag.org/content/331/6018/703.short. https://doi.org/10.1126/science.1197962

Ren, X., Han, T. X., & He, Z. (2013). *Ensemble Video Object Cut in Highly Dynamic Scenes*. 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon: IEEE. pp. 1947–1954. Retrieved from http://ieeexplore.ieee.org/document/6619098/ https://doi.org/10.1109/CVPR.2013.254

Robbins, R. J., Amaral-Zettler, L., Bik, H., Blum, S., Edwards, J., Field, D., … Wooley, J. (2012). RCN4GSC workshop report: Managing data at the Interface of biodiversity and (meta)genomics. *Standards in Genomic Sciences*, *7*(1), 159–165. https://doi.org/10.4056/sigs.3156511

Robertson, T., D¨oring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., … Desmet, P. (2014). The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PLoS ONE*, *9*(8), e102623. Retrieved from http://dx.plos.org/10.1371/journal.pone.0102623. https://doi.org/10.1371/journal.pone.0102623

Roskov, Y., Kunze, T., Paglinawan, L., Orrell, T., Nicolson, D., Culham, A., … (2013). Species 2000 & ITIS Catalogue of Life, 2013 Annual Checklist.

Sánchez-Tapia, A., de Siqueira, M. F., Lima, R. O., Barros, F. S. M., Gall, G. M., Gadelha, L. M. R., … Osthoff, C. (2018). *Model-R: A Framework for Scalable and Reproducible Ecological Niche Modeling*. High Performance Computing: 4th Latin American Conference, Carla 2017. Communications in Computer and Information Science, Buenos Aires, Argentina and Colonia, Uruguay: Springer. Vol. 796, pp. 218–232. Retrieved from http://link.springer.com/10.1007/978-3-319-73353-1_15 https://doi.org/10.1007/978-3-319-73353-1_15

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology*, *9*(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Schindel, D. E., & Cook, J. A. (2018). The next generation of natural history collections. *PLOS Biology*, *16*(7), e2006125. Retrieved from http://dx.plos.org/10.1371/journal.pbio.2006125. https://doi.org/10.1371/journal.pbio.2006125

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. Retrieved from http://www.sciencedirect.com/science/article/pii/S0893608014002135. https://doi.org/10.1016/j.neunet.2014.09.003

Schneider, F. D., Fichtmueller, D., Gossner, M. M., Güntsch, A., Jochum, M., König-Ries, B., … Simons, N. K. (2019). Towards an ecological trait-data standard. *Methods in Ecology and Evolution*, *10*(12), 2006–2019. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13288. https://doi.org/10.1111/2041-210X.13288

Schoch, C. L., Robbertse, B., Robert, V., Vu, D., Cardinali, G., Irinyi, L., … Federhen, S. (2014). Finding needles in haystacks: Linking scientific names, reference specimens and molecular data for Fungi. *Database*, *2014*, bau061. Retrieved from https://academic.oup.com/database/article-lookup/doi/10.1093/database/bau061. https://doi.org/10.1093/database/bau061

Shade, A., & Teal, T. K. (2015). Computing workflows for biologists: A roadmap. *PLOS Biology*, *13*(11), e1002303. Retrieved from http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002303. https://doi.org/10.1371/journal.pbio.1002303

Sharma, S., Ciufo, S., Starchenko, E., Darji, D., Chlumsky, L., Karsch-Mizrachi, I., & Schoch, C. L. (2018). The NCBI BioCollections database. *Database*, *2018*, bay006https://academic.oup.com/database/article/doi/10.1093/database/bay006/4904552. https://doi.org/10.1093/database/bay006

Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E., & Edwards, R. A. (2014). FOCUS: An alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ*, *2*, e425. https://doi.org/10.7717/peerj.425

Silva, G. G. Z., Green, K. T., Dutilh, B. E., & Edwards, R. A. (2016). SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics*, *32*(3), 354–361. Retrieved from http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btv584. https://doi.org/10.1093/bioinformatics/btv584

Silva, L. A. E., Siqueira, M. F., Pinto, F. D. S., Barros, F. S. M., Zimbrão, G., & Souza, J. M. (2016). Applying data mining techniques for spatial distribution analysis of plant species co-occurrences. *Expert Systems with Applications*, *43*, 250–260. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417415005783. https://doi.org/10.1016/j.eswa.2015.08.031

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Retrieved from http://arxiv.org/abs/1409.1556

Smith, V. S., Rycroft, S. D., Brake, I., Scott, B., Baker, E., Livermore, L., … Roberts, D. (2011). Scratchpads 2.0: A virtual research environment supporting scholarly collaboration, communication and data publication in biodiversity science. *ZooKeys*, *150*(150), 53–70. Retrieved from http://zookeys.pensoft.net/articles.php?id=3040. https://doi.org/10.3897/zookeys.150.2193

Soberón, J., & Peterson, A. T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical transactions of the Royal Society of London, Series B, Biological Sciences*, *359*(1444), 689–698. Retrieved from http://rstb.royalsocietypublishing.org/content/359/1444/689. https://doi.org/10.1098/rstb.2003.1439

Soberón, J. M. (2010). Niche and area of distribution modeling: A population ecology perspective. *Ecography*, *33*(1), 159–167. Retrieved from http://doi.wiley.com/10.1111/j.1600-0587.2009.06074.x. https://doi.org/10.1111/j.1600-0587.2009.06074.x

Soltis, P. S. (2017). Digitization of herbaria enables novel research. *American Journal of Botany*, *104*(9), 1281–1284. Retrieved from http://doi.wiley.com/10.3732/ajb.1700281. https://doi.org/10.3732/ajb.1700281

Souza Muñoz, M. E., Giovanni, R., Siqueira, M. F., Sutton, T., Brewer, P., Pereira, R. S., … Canhos, V. P. (2009). openModeller: A generic approach to species' potential distribution modelling. *GeoInformatica*, *15*(1), 111–135. Retrieved from http://link.springer.com/10.1007/s10707-009-0090-7. https://doi.org/10.1007/s10707-009-0090-7

Spehn, E. M., & Korner, C. (2009). *Data mining for global trends in mountain biodiversity*, Boca Raton, FL: CRC Press.

Starlinger, J., Brancotte, B., Cohen-Boulakia, S., & Leser, U. (2014). Similarity search for scientific workflows. *Proceedings of the VLDB Endowment*, *7*(12), 1143–1154. Retrieved from http://dl.acm.org/citation.cfm?doid=2732977.2732988. https://doi.org/10.14778/2732977.2732988

Stephens, C. R., Heau, J. G., González, C., Ibarra-Cerdeña, C. N., Sánchez-Cordero, V., & González-Salazar, C. (2009). Using biotic interaction networks for prediction in biodiversity and emerging diseases. *PLoS ONE*, *4*(5), e5725. Retrieved from https://dx.plos.org/10.1371/journal.pone.0005725. https://doi.org/10.1371/journal.pone.0005725

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., … Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLOS Biology*, *13*(7), e1002195. Retrieved from http://dx.plos.org/10.1371/journal.pbio.1002195. https://doi.org/10.1371/journal.pbio.1002195

Stockle, M., & Hebert, P. (2008). Barcode of life. *Scientific American* Retrieved from http://www.nature.com/scientificamerican/journal/v299/n4/full/scientificamerican1008-82.html, *299*, 88.

Stocks, K. I., Stout, N. J., & Shank, T. M. (2016). Information management strategies for deep-sea biology. In M. R. Clark, M. Consalvey, & A. A. Rowden (Eds.), *Biological sampling in the deep sea* (pp. 368–385). Hoboken, NJ: Wiley Blackwell.

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE*, *8*(6), e67111. Retrieved from https://dx.plos.org/10.1371/journal.pone.0067111. https://doi.org/10.1371/journal.pone.0067111

Strasser, C., Abrams, S., & Cruse, P. (2014). DMPTool 2: Expanding functionality for better data management planning. *International Journal of Digital Curation*, *9*(1), 324–330. https://doi.org/10.2218/ijdc.v9i1.319

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/J.BIOCON.2013.11.003

Talbert, C., Talbert, M., Morisette, J., & Koop, D. (2013). Data management challenges in species distribution modeling. *IEEE Bulletin of the Technical Committee on Data Engineering*, *36*(4), 31–40.

Tan, P.-N., Kumar, V., & Srivastava, J. (2002). *Selecting the Right Interestingness Measure for Association Patterns*. Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining—KDD'02. ACM Press, New York, NY. p. 32. Retrieved from http://portal.acm.org/citation.cfm?doid=775047.775053 https://doi.org/10.1145/775047.775053

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology & Evolution*, *18*(2), 70–74. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169534702000411. https://doi.org/10.1016/S0169-5347(02)00041-1

Thébault, E. (2013). Identifying compartments in presence-absence matrices and bipartite networks: Insights into modularity measures. *Journal of Biogeography*, *40*(4), 759–768. https://doi.org/10.1111/jbi.12015

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., … Williams, S. E. (2004). Extinction risk from climate change. *Nature*, *427*(6970), 145–148. Retrieved from http://www.nature.com/doifinder/10.1038/nature02121. https://doi.org/10.1038/nature02121

Tulloch, A. I. T., Chadès, I., Dujardin, Y., Westgate, M. J., Lane, P. W., & Lindenmayer, D. (2016). Dynamic species co-occurrence networks require dynamic biodiversity surrogates. *Ecography*, *39*(12), 1185–1196. Retrieved from http://doi.wiley.com/10.1111/ecog.02143. https://doi.org/10.1111/ecog.02143

Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., & Steininger, M. (2003). Remote sensing for biodiversity science and conservation. *Trends in Ecology & Evolution*, *18*(6), 306–314. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169534703000703. https://doi.org/10.1016/S0169-5347(03)00070-3

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & De Clerck, O. (2012). Bio-ORACLE: A global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, *21*(2), 272–281. Retrieved from http://doi.wiley.com/10.1111/j.1466-8238.2011.00656.x. https://doi.org/10.1111/j.1466-8238.2011.00656.x

Ulloa, C. U., Acevedo-rodríguez, P., Beck, S., Belgrano, M. J., Bernal, R., Berry, P. E., … Jørgensen, P. M. (2017). An integrated assessment of the vascular plant species of the Americas. *Science*, *358*(6370), 1–5. https://doi.org/10.1126/science.aao0398

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11), 1084–1091. http://doi.wiley.com/10.1111/j.1600-0587.2013.00441.x. https://doi.org/10.1111/j.1600–0587.2013.00441.x

Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., & Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLOS ONE*, *12*(6), e0178731. Retrieved from http://dx.plos.org/10.1371/journal.pone.0178731. https://doi.org/10.1371/journal.pone.0178731

Vicario, S., Balech, B., Donvito, G., Notarangelo, P., & Pesole, G. (2012). *The BioVel Project: Robust phylogenetic workflows running on the GRID*. Vol. 18. No. B. Retrieved from http://journaldev.embnet.org/index.php/embnetjournal/article/view/557

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0148296317305441. https://doi.org/10.1016/j.jbusres.2017.12.043

Wagner, R. A., & Lowrance, R. (1975). An extension of the string-to-string correction problem. *Journal of the ACM*, *22*(2), 177–183. Retrieved from http://portal.acm.org/citation.cfm?doid=321879.321880. https://doi.org/10.1145/321879.321880

Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., … Wooley, J. (2014). Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE*, *9*(3), e89606. Retrieved from http://dx.plos.org/10.1371/journal.pone.0089606. https://doi.org/10.1371/journal.pone.0089606

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One*, *7*(1), e29715. Retrieved from http://dx.plos.org/10.1371/journal.pone.0029715. https://doi.org/10.1371/journal.pone.0029715

Wiens, J. A., Stralberg, D., Jongsomjit, D., Howell, C. A., & Snyder, M. A. (2009). Niches, models, and climate change: Assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences*, *106*(Supplement 2), 19729–19736. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.0901639106. https://doi.org/10.1073/pnas.0901639106

Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K. P., … Meyer, F. (2016). The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Research*, *44*(D1), D590–D594. Retrieved from http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1322. https://doi.org/10.1093/nar/gkv1322

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. Retrieved from http://www.nature.com/articles/sdata201618. https://doi.org/10.1038/sdata.2016.18

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., … Wilson, P. (2014). Best practices for scientific computing. *PLoS Biology*, *12*(1), e1001745. Retrieved from http://dx.plos.org/10.1371/journal.pbio.1001745. https://doi.org/10.1371/journal.pbio.1001745

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, *6*(2), e1000667. Retrieved from http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000667{#}pcbi-1000667-g004. https://doi.org/10.1371/journal.pcbi.1000667

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. Retrieved from http://link.springer.com/10.1007/s10115-007-0114-2. https://doi.org/10.1007/s10115-007-0114-2

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., … Culham, A. (2007). How global is the global biodiversity information facility? *PLoS One*, *2*(11), e1124. Retrieved from http://dx.plos.org/10.1371/journal.pone.0001124. https://doi.org/10.1371/journal.pone.0001124

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), 415–420. Retrieved from https://doi.org/10.1038/nbt.1823. https://doi.org/10.1038/nbt.1823

Zaharia, M., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., … Venkataraman, S. (2016). Apache spark: A unified engine for big data processing. *Communications of the ACM*, *59*(11), 56–65. Retrieved from http://dl.acm.org/citation.cfm?doid=3013530.2934664. https://doi.org/10.1145/2934664

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., B¨ackstr¨om, D., Juzokaite, L., Vancaester, E., … Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), 353–358. Retrieved from http://www.nature.com/articles/nature21031. https://doi.org/10.1038/nature21031

How to cite this article: Gadelha LMR Jr, de Siracusa PC, Dalcin EC, et al. A survey of biodiversity informatics: Concepts, practices, and challenges. *WIREs Data Mining Knowl Discov*. 2021;11:e1394. https://doi.org/10.1002/widm.1394